Psychological Bulletin

CONTENTS

Published Bimonthly by the American Psychological Association

Vol. [illegible] No. 2 March, 1957

The Psychological Bulletin contains evaluative reviews of research literature and articles on research methodology in psychology. *This journal does not publish reports of original research or original theoretical articles.*

Manuscripts should be sent to Wayne Dennis, Department of Psychology, Brooklyn College, Brooklyn 10, New York.

Preparation of articles for publication. Authors are [illegible] advised to follow the general directions given in the "Publication Manual of the American Psychological Association" (*Psychological Bulletin*, 1952, 49 [No. 4, Part 2], 389–449). Special attention should be given to the section on the preparation of the references (pp. 432–440), since this is a particular source of difficulty in long reviews of research literature. *All copy must be double spaced, including the references.* All manuscripts should be submitted *in duplicate.* Original figures are prepared for publication; duplicate figures may be photographic or pencil-drawn copies. Authors are cautioned to retain a copy of the manuscript to guard against loss in the mail.

Reprints. Fifty free offprints are given to contributors of articles and notes. Authors of early publication articles receive no gratis offprints.

Communications—including subscriptions, orders of back issues, and changes of address—should be addressed to the American Psychological Association, 1333 Sixteenth Street N.W., Washington 6, D. C. Address changes must reach the Subscription Office by the 10th of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

Annual subscription: $8.00 (Foreign $8.50). Single copies, $1.50.

PUBLISHED BIMONTHLY BY:

THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.

Menasha, Wisconsin,
and 1333 Sixteenth Street N.W., Washington 6, D.C.

Entered as second class mail matter at the post office at Washington, D.C., under the Act of March 3, 1879. Additional entry at the post office at Menasha, Wisconsin. Acceptance for mailing at special rate of postage under the provisions of Sec. 34-40 Par. (D) provided for in Section 538, act of [illegible], 19[illegible], authorized [illegible] 1947. Printed in U.S.A.

VOL. 54, NO. 2 MARCH, 1957

Psychological Bulletin

MEASUREMENT OF REPRODUCIBILITY[1]

BENJAMIN W. WHITE
Lincoln Laboratory, Massachusetts Institute of Technology
AND ELI SALTZ
Air Force Personnel and Training Research Center, Chanute Air Force Base, Illinois

Much of our knowledge of human behavior is based upon data obtained through the administration of multiple-choice tests to groups of subjects. Such instruments are used in many ways: selection, attitude measurement, ability measurement, and clinical diagnosis, to name only a few. Particularly since the publication of Guttman's model for measuring a test's reproducibility (7), there has been increasing concern over one aspect of the responses of groups of subjects to groups of items—the extent to which the patterns of subjects' responses can be predicted from their total scores. While these considerations have been of great interest to social and clinical psychologists, they have also proved pertinent to constructors of ability tests. It is the purpose of this article (*a*) to examine some of the techniques which have been devised to assess a test's "reproducibility," "homogeneity," or internal consistency, (*b*) to evaluate these techniques against certain criteria, and (*c*) to suggest possible logical relationships of these techniques to the concept of reliability.

[1] The opinions and conclusions contained in this article are those of the authors. They are not to be construed as reflecting the views or endorsement of the Department of the Air Force.

In the ensuing discussion the word *test* will be used to describe any technique whereby two or more subjects respond to two or more stimuli in such a way that the responses of all subjects to each item can be dichotomized. It is assumed that every subject responds to every such item. It is further assumed that the experimenter assigns a value of unity to all responses on one side of the dichotomy and a value of zero to the rest. A "total score" for a subject is computed by adding the weights assigned to his responses thus dichotomized. With this system, a subject's total score is the number of responses he has made which fall into the unity-weighted class.

Often such scores are presumed to yield an ordering of the subjects on some hypothetical linear continuum, ability, or trait. For some time social scientists have been aware that this process of assigning a simple order to people on the basis of their responses to a number of test items is a legitimate representation of their test behavior only when their responses possess certain characteristics. There are many ways of stating this, but for the purposes of this discussion, it will be most convenient to use the following: a total score, computed by counting the number of test responses which have been classi-

fied in one of two ways, will yield a perfect mapping of the entire pattern of responses of all subjects when, and only when, the interitem covariances are maximal.

For purposes of illustration, consider a six-item test. On such a test, total scores can take seven possible values from 0 to 6. When interitem covariance is maximal, there is only one way in which a subject can make any given total score. Naturally he can make a total score of 0 only by "failing" all six items, and a score of 6 only by "passing" all six items. He can make a score of 1 only by passing the easiest item. By "easiest" is meant the item which was passed by more subjects than any other. Similarly he can make a score of 2 only by passing the two easiest items. In other words, given the information that the interitem covariances are maximal, the order of difficulty of the items, and a subject's total score, one can tell exactly which items the subject got wrong and right. On such a test there are only seven ways in which people respond to the items, and each of these corresponds with one of the seven possible total scores.

At the other extreme, consider a six-item test whose items are independent, i.e., exhibit zero covariances. Such a test could yield 2^6 or 64 different response patterns. There would be 15 different ways in which a person could get a total score of 2, for example. In this situation, given knowledge of the total score, the order of difficulty of the items, and the fact of zero covariance between items, one would not be able to reconstruct a subject's pattern of responses to the test, unless the total score happened to be 0 or 6. Representation of the test behavior of the subjects with the conventional total score would result in a considerable loss of information.

Various indices have been developed which will permit the tester to ascertain the degree to which the total scores of a given test yield a complete mapping of the responses of all subjects to all the items (reproducibility). These indices differ not only in their computational formulas, but in their underlying assumptions, though all start with the same primary data: the dichotomized responses of a group of subjects to a group of test items. Four criteria are suggested against which each index may be evaluated.

1. *Does it yield a theoretical maximum value which is the same for any test?*

2. *Does it yield a theoretical minimum value which is the same for any test?*

3. *Does it permit evaluation of the null hypothesis that the obtained reproducibility index is not significantly different from chance?*

4. *Does it permit evaluation of each item in the test as well as of the test as a whole?*

The rationales for these criteria are reasonably straightforward. If maximum or minimum possible values differ from test to test, it is difficult to evaluate one test against another. For example, two tests having reproducibility quotients of .90 are differently evaluated when it is discovered that the minimum theoretical reproducibility of one is .60, and of the other .90. If the quotient does not have a known sampling distribution, there is the possibility that the obtained quotient does not differ significantly from chance. And finally, if the items can not be evaluated it is difficult to improve reproducibility by omission or inclusion of specific items.

In the light of these criteria, we propose to discuss several techniques which have been devised to yield an

TABLE 1

RESPONSES OF TEN SUBJECTS TO A SIX-ITEM TEST WHERE ROWS AND COLUMNS ARE UNORDERED

Subject	Item						Total Score
	1	2	3	4	5	6	
A	0	0	0	1	0	0	1
B	1	0	1	1	1	1	5
C	0	0	1	0	0	0	1
D	1	1	1	0	1	1	5
E	1	0	1	1	1	0	4
F	0	0	1	0	1	0	2
G	0	1	1	0	0	0	2
H	1	0	1	0	1	0	3
I	1	0	1	0	1	1	4
J	1	0	0	1	1	0	3
Item Difficulty	6	2	8	4	7	3	

index of reproducibility. In order to demonstrate the computations involved in each technique, we shall use the responses of ten subjects to a six-item test, illustrated in Table 1.

In this matrix the rows represent subjects and the columns test items. The marginal entries at the bottom of the matrix indicate the number of subjects who "passed" a given item, and the marginals in the last column of the matrix represent the number of items each subject "passed."

GUTTMAN'S REPRODUCIBILITY

Guttman (7) originated the term reproducibility. The term means essentially the degree to which one can reproduce a subject's entire response pattern from a knowledge of his total score and the order of difficulty of the items. Originally Guttman's technique of obtaining the index of reproducibility involved mechanical operations on a matrix of N subjects and K test items similar to Table 1. A device, the scalogram board, permits interchange of rows and columns of this matrix in a particular manner so that the unity entries are maximally concentrated above the main diagonal of the matrix. Such rearrangement of the response matrix in Table 1 is shown in Table 2.

It should be noted that if there are any ties in total score or in the number of subjects passing items, the arrangement of the matrix may not be unique. In this example the order of the columns is unique since there are no ties in the number of subjects passing items, but the order of rows is not, since there are two subjects at each total score level. In such

TABLE 2

RESPONSES OF TEN SUBJECTS TO A SIX-ITEM TEST WHERE ROWS AND COLUMNS ARE ORDERED

Subject	Item						Total Score
	2	6	4	1	5	3	
D	1	1	0	1	1	1	5
E	0	1	1	1	1	1	5
B	0	0	1	1	1	1	4
I	0	1	0	1	1	1	4
H	0	0	0	1	1	1	3
J	0	0	1	1	1	0	3
G	1	0	0	0	0	1	2
F	0	0	0	0	1	1	2
C	0	0	0	0	0	1	1
A	0	0	1	0	0	0	1
Item Difficulty	2	3	4	6	7	8	

TABLE 3

JACKSON'S METHOD OF COMPUTING REPRODUCIBILITY (R), MINIMUM REPRODUCIBILITY (MR), AND PLUS PERCENTAGE RATIO (PPR)

Subject	Item											
	2		6		4		1		5		3	
	+	−	+	−	+	−	+	−	+	−	+	−
D	1		1			(0)	1		1		1	
E		0	1		1		1		1		1	
B		0		0	1		1		1		1	
I		0	(1)			0	1		1		1	
H		0		0		0	1		1		1	
J		0		0	(1)		1		1			(0)
G	(1)			0		0		0		(0)	1	
F		0		0		0		0	1		1	
C		0		0		0		0		0	1	
A		0		0	(1)			0		0		0
# right (P)	2		3		4		6		7		8	
# wrong (Q)		8		7		6		4		3		2
Errors	1		1		3		0		1		1	
R_i	.90		.90		.70		1.00		.90		.90	
MR_i	.80		.70		.60		.60		.70		.80	
PP_i	.10		.20		.10		.40		.20		.10	
PPR_i	.50		.67		.25		1.00		.67		.50	

Note.—Rights are listed under +. Wrongs are listed under −.
Total errors = 7; R_i = 88%; MR_i = 70%; PP_i = 18%; PPR_i = .61.

cases further permutations of rows and columns are made until errors are minimized. The index of reproducibility is a function of the number of errors, i.e., unity entries which are below the diagonal and zero entries which are above it. This diagonal is not necessarily exactly coincident with the main diagonal, and Guttman has several rules to be followed in its determination. Since Guttman's original procedure is unwieldy, we shall in this illustration use a procedure developed by Jackson (**10**) for arriving at cutting points for each item. For all practical purposes, Jackson's R_i quotient is identical with Guttman's.[2] Jackson's method is illustrated in Table 3 above.

[2] It should be noted that many people have suggested modifications in the calculations of Guttman's R_i (**3, 6, 11, 17, 18**). These refinements of procedure are, by and large, identical in their logical properties with

This is the same matrix shown in Table 2, except that the unity and zero entries under each item have been placed in separate columns. In order to draw cutting points, one simply draws a line across each column at the place where the number of zero entries above the line and the number of unit entries below the line (errors) are minimized. These cutting points are seen as descending steps in the table. In the first column there is one entry of unity which falls below the cutting line and this has been put in parentheses. If the cutting line had been drawn directly below this unity entry, the five zero entries above it would be counted as errors and put in parentheses. In this illustration there is a unique cutting point for five of the six items, i.e., a line which yields an absolute minimum number of errors. In Item 5 however, the line could be either where it is drawn, or two rows higher. Either solution yields 1 error. The lower one was chosen because it yielded an additional cutting point for the scale,[3] whereas the higher cutting point would have been identical for that of Item 1.

After the cutting points have been assigned, the errors in each column are counted. From these it is possible to compute the reproducibility for each item (R_i) by dividing the number of errors (E) by the number of subjects (N) and subtracting the quotient from 1.

$$R_i = 1 - \frac{E}{N}. \qquad [1]$$

The reproducibility for the entire test (R_t) may be computed by summing the errors for all items

$$\left(\sum_{i=1}^{k} E\right)$$

dividing this by the number of subjects (n) times the number of items (k), and subtracting the quotient from 1.

$$R_t = 1 - \frac{\sum_{i=1}^{k} E}{NK}. \qquad [2]$$

For this example, the reproducibility of the test is 88.3 per cent, somewhat below the 90 per cent figure which Guttman uses as one criterion of scalability.

The Guttman index of reproducibility meets our first criterion in that it has an absolute maximum of 100 per cent for any test with more than one item, and our fourth criterion in that one can compute the index for each item as well as for the test as a whole. However, it suffers a serious shortcoming in having no unique minimal value. As Jackson (10)

Guttman's quotient, and were so intended by their authors. Consequently no space is given to them in this article.

[3] In Jackson's method, the cutting points are used to determine minimum number of errors. Once the minimum number of errors has been determined the exact locations of the cutting points no longer enter into the computation of reproducibility. Consequently, for Jackson's method it doesn't matter which of the two cutting points is used for Item 5, since both result in one error. However, Guttman's original procedure made use of the specific cutting point used. Guttman assigned the cutting points to the row marginals (the total scores) and then rescored every *S* on the basis of the cutting points. All *S*s below the lowest cutting point would be scored as having failed all the items. In Table 3, for example, Item 3 has the lowest cutting point; subject A is below this cutting point and so he would be rescored as having failed all the items. All *S*s between the first and second cutting points would be rescored as having passed one item. And so forth. The Guttman reproducibility index indicates the percentage of actual reproducibility as compared with the maximum reproducibility obtained by these rescoring processes. Therefore, if two items are given the same cutting point, the number of different classes or "cutting point scores" will be decreased—that is, the number of discriminations made by the scale is diminished.

and others (**1, 2, 13, 14**) have pointed out, the index of reproducibility is drastically affected by the difficulty levels of the items in a test. The reason for this is that the difficulty of an item (percentage of persons passing) places a limit on the likelihood of an error: passing a difficult item, and failing an easy one. The reproducibility figure can approach its absolute lower limit of 50 per cent only when all the items have a difficulty level of 50 per cent, a trivial case in which 100 per cent reproducibility could be obtained only if one-half the subjects passed all the items while the other half failed all the items. With even slight departures from this strict condition, the lower limit of the reproducibility index rises sharply. In our illustrative example minimum reproducibility is 70 per cent. This fact makes it exceedingly difficult to evaluate an obtained index of reproducibility. With short scales and wide spread in item difficulties, Guttman's figure of 90 per cent may on occasion be very little higher than the minimum reproducibility of the scale.

Jackson's Plus Percentage Ratio (*PPR*)

In order to circumvent this drawback of Guttman's reproducibility index, Jackson (**10**) has developed another statistic which he calls the Plus Percentage Ratio (*PPR*). Unlike the Guttman index, *PPR* has the same absolute minimum for all tests. Referring again to Table 3, note the minimum reproducibility figures in the row labelled MR_i. Here the minimum reproducibility figure for each item (MR_i) is obtained by dividing the number of subjects who got a given item right (# right), or wrong (# wrong), whichever figure is the larger, by the number of subjects (N).

$$MR_i = \frac{\#\text{ rights or }\#\text{ wrongs (whichever is larger)}}{N}. \quad [3]$$

The minimum reproducibility for the entire test (MR_t) is computed by taking for each item the number of rights

$$\left(\sum_{i=1}^{k} \#\text{ rights}\right)$$

or the number of wrongs

$$\left(\sum_{i=1}^{k} \#\text{ wrongs}\right),$$

whichever number is larger, summing the numbers so obtained over all items and dividing this sum by the product of the number of items (K) and the number of subjects (N).

$$MR_t = \frac{\sum_{i=1}^{k} \#\text{ rights, or }\#\text{ wrongs (whichever is larger)}}{KN}. \quad [4]$$

In the next to the last row of Table 3, the "Plus $\%_i$" (PP_i) figures are listed. Here the differences between the obtained reproducibility and the minimum reproducibility ($R_i - MR_i$) for each item are entered. These figures indicate how much better obtained reproducibility is than the minimum for that item. In the last row, the "Plus % Ratios" (PPR_i) are entered for each item. These figures may be obtained by dividing the Plus % figure for a given item by one minus the minimum reproducibility (MR_i) for that item.

$$PPR_i = \frac{R_i - MR_i}{1 - MR_i}. \quad [5]$$

The Plus Percentage Ratio for the total test (PPR_t) is similarly computed by dividing the difference between R_t and MR_t by one minus MR_t.

$$PPR_t = \frac{R_t - MR_t}{1 - MR_t}. \qquad [6]$$

The Plus % Ratio has a distinct advantage over the index of reproducibility in that it has both an absolute maximum of one and an absolute minimum of zero for any test of more than one item. For the test illustrated here the PPR_t is .61. As Jackson points out, testers should be prepared for the fact that this index will almost inevitably be lower than the Guttman index of reproducibility, often considerably lower. The index has not often been used on well-known tests, so it is difficult to say what an acceptable level should be. Jackson tentatively suggests 70 per cent. It remains to be seen whether this figure is a reasonable one in mental testing or attitude scaling. The *PPR* in any event has much to recommend it since it circumvents one of the most serious criticisms which has been leveled at Guttman's reproducibility index.

Loevinger's Index of Homogeneity (H)

Homogeneity of a Test (H_t)

A rather different approach to the measurement of the reproducibility of mental tests has been put forth by Loevinger, who uses the following as a definition of homogeneity (**13**, p. 29).

> The definitions of perfectly homogeneous and perfectly heterogeneous tests can be restated in terms of probability. In a perfectly homogeneous test, when the items are arranged in the order of increasing difficulty, if any item is known to be passed, the probability is unity of passing all previous items. In a perfectly heterogeneous test, the probability of an individual passing a given item A is the same whether or not he is known already to have passed another item B.

It can be seen that this definition comes quite close to the Guttman notion of reproducibility, and in fact the perfectly reproducible and the perfectly homogeneous test are identical.

With the test items arranged in order of increasing difficulty, Loevinger computes the quantity S by finding, for *all pairs of items*, the proportion of subjects who have passed both items (P_{ij}). From this is subtracted the theoretical proportion who would have passed both items had they been independent (P_iP_j). These differences are summed over the $k(k-1)$ pairs of items (i.e., each item is paired with every other item in the test).

$$S = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} P_{ij} - P_iP_j. \qquad [7]$$

For a test made up of completely independent items, S would have a value of zero. S does not have an upper limit of unity when the test is perfectly homogeneous. The upper limit is fixed by the proportion of subjects passing the more difficult item in each pair (P_j).

$$S_{max} = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} P_j - P_iP_j. \qquad [8]$$

The homogeneity of a test (H_t) is then given by the ratio of these two quantities

$$H_t = \frac{S}{S_{max}}. \qquad [9]$$

This procedure is exactly analogous to that used by Jackson in computing the Plus Percentage Ratio. This can be seen more easily if Loevinger's equation is rewritten as follows:

$$H_t = \frac{S}{S_{max}} = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} (1 - P_{ij}) - (1 - P_iP_j)}{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} 1 - (1 - P_iP_j)}. \qquad [10]$$

The first term in the parentheses of the numerator $(1-P_{ij})$ indicates the proportion of subjects passing a harder item *and* failing an easier one subtracted from unity. This is very like the reproducibility coefficient which is given by the proportion of errors subtracted from unity. The second term in the numerator $(1-P_iP_j)$ is the product of the proportion of subjects passing the harder item and the proportion failing the easy item, this product then subtracted from unity. The quantity $(1-P_iP_j)$ is analogous to Jackson's minimum reproducibility. The denominator is seen to be the difference between unity (perfect reproducibility) and minimum reproducibility. The two methods differ only in the procedure for counting errors. Loevinger's technique involves the equivalent of an examination of all pairs of items $i \neq j$ and counting every occasion upon which the harder item is passed and the easier item failed. In the illustrative example, such a tabulation yields a total of 13 errors, whereas Jackson's error count is 7. This is the reason that Loevinger's H_i will usually be lower than Jackson's PPR_i. The former is .23, and the latter .61. In Jackson's system for counting errors, a deviant response is counted only once no matter where it occurs in the response pattern. For example, if items are arranged in order of decreasing difficulty, a response pattern of (1, 0, 0, 0) would be credited with one error, while in Loevinger's system, since the passed item was the hardest of the four, there would be three errors. The two methods also have somewhat different ways of computing minimum reproducibility, Jackson's yielding a figure of .70, and Loevinger's .72.

Loevinger points out that her formula for H_i is equivalent to

$$H = \frac{\sigma^2_x - \sigma^2_{het}}{\sigma^2_{hom} - \sigma^2_{het}}, \qquad 11$$

where all the variances refer to total raw scores. The first term in the numerator (σ^2_x) is the variance of the obtained scores, the second numerator term (σ^2_{het}) is the variance of the total scores which would be obtained from items of the same difficulties which were completely independent, and the first term in the denominator (σ^2_{hom}) is the variance in total scores which would be obtained if the same items were perfectly correlated. The raw score variance of a test made up entirely of independent items is the familiar

$$\sum_{i=1}^{k} pq,$$

or the sum of the item variances. The raw score variance of a test made up wholly of perfectly correlated items is given by

$$\sigma^2_{hom} = \sum_{i=1}^{k} P_iQ_i + 2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} P_j - P_iP_j. \quad [12]$$

The first term on the right of this equation is the item variance employed above, and the second term is two times a sum which is seen to be identical to S_{max}.

This relationship is interesting since it shows that total score variance increases with reproducibility, being at a minimum when the item covariances are zero, and reaching an upper limit when item covariances are maximal.

Both Loevinger's H_i and Jackson's *PPR* have the advantage of being uninfluenced by the distribution of item difficulties which makes them preferable to the Guttman reproducibility index when it is given without further information. The procedures are objective and can be reduced to routine computations. When "errors" occur mainly on item pairs which are

close together in difficulty level, the two procedures should yield practically identical indices, but if there are "errors" which occur in item pairs which are widely different in difficulty level, Loevinger's H_t will be lower than Jackson's PPR_t. Loevinger's technique has the aesthetic advantage of making full use of the information contained in the response matrix, but the practical drawback of being tedious to compute when the number of items is large since $k(k-1)$ cross breaks have to be made to compute the P_{ij}s. However, Jackson's method is also laborious since it requires an initial posting of the entire response matrix.

The sampling distribution of H_t is unknown, and Loevinger advises that it should not be used as an estimate of homogeneity unless the sample of subjects exceeds 100.

Homogeneity of an Item with a Test (H_{it})

Loevinger's H_t yields an index for the test as a whole, but does not provide an index of the homogeneity of each item with the test. For this purpose, she suggests another index, (H_{it}), the logic of which is the same as that employed in H_t. In a perfectly homogeneous test, subjects passing a given item should have higher total scores than those failing the item. The starting point is a formula developed by Long (**15**).

$$\text{Long's Index} = 1 - \frac{2\sum \text{"passes" below "fails"}}{PQ}. \quad [13]$$

In 13, the numerator is two times the number of subjects passing a given item who have total scores lower than those of subjects who failed the same item, and the denominator is the product of the number of passes on the item (P) and the number of fails (Q). Loevinger points out that difficulties arise when two subjects have identical total scores, one of whom has failed the item and the other has passed it. There is also the question of whether the response to the item should in this computation be included in the total score. In order to circumvent these difficulties with Long's index, Loevinger proposes the modification

$$H_{it} = 1 - \frac{2\sum \text{"passes" below or tied with "fails"}}{PQ - \sum \text{"passes" one above "fails"}}. \quad [14]$$

It is clear that this index can take values from minus to plus unity, but it is not clear that a zero value is obtained when there is no relation between an item and the total test. The sampling properties of the index are unknown and will have to be investigated to establish the value to be expected for a chance relation. The obtained H_{it} values for the illustrative test may be seen in the last column of Table 6.

GREEN'S SUMMARY STATISTICS METHOD (I)

Green (**4, 5**) has recently developed a method for computing an index of consistency for a test (I) which has all the advantages of Jackson's PPR, and Loevinger's H_t, plus greater ease of computation. Like Jackson's PPR, I is given by

$$I = \frac{Rep - Rep_{ind}}{1.00 - Rep_{ind}}, \quad [15]$$

where Rep is the obtained reproducibility of the test, Rep_{ind} is the reproducibility which would be obtained with the same set of item difficulties and complete independence between items, and 1.00 is perfect reproducibility.

Green's method of computing errors is the same as that employed in

10 above, except that the summation is not over all pairs of items, $i \neq j$, but only over those item pairs whose members are adjacent in difficulty level. Green's reproducibility is given by

$$Rep = 1 - \frac{1}{NK}\sum_{i=1}^{k-1} n_{i,i+1} - \frac{1}{NK}\sum_{i=2}^{k-2} n_{\overline{i-1},i,i+1,i+2}, \quad [16]$$

where N is the number of subjects, K the number of items. Items are ranked in order of difficulty, the most difficult item receiving rank k, and the easiest item rank 1. The quantity $n_{i,i+1}$ is the number of subjects who both fail the ith item and pass the next most difficult item $(i+1)$. There will be $k-1$ such item pairs. The last quantity, $n_{\overline{i-1},i,i+1,i+2}$ is the number of subjects who have failed both item $i-1$ and i and passed both item $i+1$ and $i+2$. There will be $k-3$ such terms in this summation.

The reproducibility that would be expected if the items had their observed difficulties, but were mutually independent is given by

$$Rep_{ind} = 1 - \frac{1}{N^2K}\sum_{i=1}^{k-1} n_i n_{i+1} - \frac{1}{N^4K}\sum_{i=2}^{k-2} n_i n_{i+1} n_{i+2} n_{\overline{i-1}}. \quad [17]$$

These values for *Rep* and Rep_{ind} are then put in 15 to obtain I, which will be unity for a perfectly reproducible test and zero for a test whose items are completely independent. Green suggests that I should be .50 for a test before its items can be considered scalable. Since this method makes only a partial count of the "errors" in a response matrix, it produces a slight overestimate of reproducibility. In one empirical investigation (**5**) it was found that the average discrepancy between Green's reproducibility and the exact reproducibility of ten scales was .002.

Following a suggestion of Guttman (**7**), Green furnishes an approximation to the standard error of *Rep*.

$$\sigma_{Rep} \approx \sqrt{\frac{(1-Rep)(Rep)}{NK}}. \quad [18]$$

With this standard error it is possible to ascertain whether an obtained *Rep* is significantly larger than Rep_{ind}. Green warns, however, that when such a test yields borderline significance, one should be cautious in interpretation since both *Rep* and σ_{Rep} are approximations. A high significance level does not necessarily indicate that the items are homogeneous, merely that the item intercorrelations are significantly greater than zero.

For the illustrative test, the computation of *Rep*, Rep_{ind}, and I are shown in Table 4.

The obtained *Rep* is .917, as compared with Jackson's .88, and .78 by Formula 10. The index of consistency (I) is seen to be .41, as compared with Jackson's *PPR* of .61, and Loevinger's H_t of .23.

The Phi Coefficient (ϕ_{it})[4]

A measure of item reproducibility can be derived from the phi coefficient. This measure has the advantages of an absolute maximum of 1.00, an absolute minimum of 0.00, a known sampling distribution, and direct relationship to conventional test construction procedure.

The logic behind the procedure is simple. Take as an example an item which 30 per cent of the subjects pass and 70 per cent fail. If the item is perfectly reproducible in a perfectly reproducible test, the 30 per cent of the subjects with the highest total

[4] The writers find that Cronbach (**1**, p. 324) has anticipated them in this suggested manner of estimating reproducibility.

TABLE 4

GREEN'S METHOD OF COMPUTING REPRODUCIBILITY (*Rep*), CHANCE REPRODUCIBILITY (Rep_{ind}), AND INDEX OF CONSISTENCY (*I*)

Subjects	Items						Total Scores
	2	6	4	1	5	3	
D	1	1	0	1	1	1	5
E	0	1	1	1	1	1	5
B	0	0	1	1	1	1	4
I	0	1	0	1	1	1	4
H	0	0	0	1	1	1	3
J	0	0	1	1	1	0	3
G	1	0	0	0	0	1	2
F	0	0	0	0	1	1	2
C	0	0	0	0	0	1	1
A	0	0	1	0	0	0	1
Rank Order of Difficulty	6	5	4	3	2	1	

n_i	2	3	4	6	7	8
n_j	8	7	6	4	3	2
$n_{i,i+1}$	–	1	2	1	0	1
$n_{i-1,i,i+1,i+2}$	–	–	0	0	0	–

$$Rep = 1 - \frac{1}{(10)(6)}(1+0+1+2+1) - \frac{1}{(10)(6)}(0+0+0) = .917$$

$$Rep_{ind} = \frac{1}{(10^2)(6)}(7\cdot2+6\cdot3+4\cdot4+3\cdot6+2\cdot7) - \frac{1}{10^4(6)}(4\cdot6\cdot3\cdot2+3\cdot4\cdot4\cdot3+2\cdot3\cdot6\cdot4) = .860$$

$$I = \frac{Rep - Rep_{ind}}{1.00 - Rep_{ind}} = \frac{.916 - .860}{1.000 - .860} = .407$$

scores should all pass the item; the 70 per cent with the lowest total scores should all fail the item. Subjects can easily be ranked on total score and this distribution cut in the same ratio as the pass-fail ratio on any particular item being evaluated. It is then simple to determine the number of persons high on total score who pass the item, the number of high persons who fail the item, the number of low persons who fail the item and who pass the item. The data may be put in a fourfold table as in Table 5.

TABLE 5

ITEM-TOTAL SCORE PHI COEFFICIENT (ϕ_{it})

		Total Score*		Total
		Low	High	
Item Score	Pass Item *i*	*A*	*B*	*A*+*B*
	Fail Item *i*	*C*	*D*	*C*+*D*
	Total	*A*+*C*	*B*+*D*	*N*

* Total score distribution is broken so that number of subjects in low group is equal to number failing item *i*: $(C+D = A+C)$.

Obviously, one has only to determine the marginal sums (which are determined by the pass-fail ratio of the item) and one of the cell frequencies, since the rest can be computed by subtraction from the marginals.

Splitting subjects on the basis of total score in the same ratio as the pass-fail split on an item may produce a problem if several subjects are tied for total score across the cutting points. The tied subjects should be randomly distributed between the high and low groups so that the total scores are split in the same ratio as the pass-fail ratio. Take as a simple example the case in which 100 subjects have answered a questionnaire in such a manner that the pass-fail ratio on a particular item is 30/70. To evaluate this item, the subjects must be split on total score so that the highest 30 per cent constitute one group and the lowest 70 per cent constitute the second group. If three persons are tied for rank 30 in total score, two will be arbitrarily considered ranks 29 and 30 respectively,

and will be placed in the high group. The third person will be assigned rank 31, and, despite the fact that his score is the same as that of two subjects in the high group, he will be placed in the low group. If the total number of subjects is reasonably large, and if the number of subjects having the *critical tied score* is not a large percentage of the total number of subjects, this will not distort the resulting phi.

Since the marginals for the total score have been determined in a manner that forces them to be equal for the marginal for the particular item the usual phi formula can be simplified to

$$\phi_{it}=\frac{BC-AD}{(A+B)(C+D)}, \qquad [19]$$

where the quantities A, B, C, and D correspond to cell entries in Table 5 above.

The null hypothesis for such a phi coefficient is, in every case, that the obtained phi is not significantly greater than zero. This can be tested by a chi square or a Fisher exact test on the fourfold table.

If the investigator desires to "purify" his test, he must choose a cutting point and select all the items with phi coefficients above this cutting point to constitute his reproducible scale. New total scores can then be computed on the basis of the selected items, and phi coefficients recalculated to give an estimate of the reproducibility of the new scale. The coefficients for some of the items not included in the new total score may be so high that these items can be included in the scale, while those for some of the included items may drop to a level which makes it advisable to exclude them.

Unlike some indices of reproducibility, this index is not affected by extremes of item difficulty. This is true because phi is not an index of the frequency in one cell, but is determined by the intercorrelation between cells. The method has a disadvantage, a purely aesthetic one, but one that may prejudice some workers against it; the phi coefficients so obtained are not likely to yield many values in the .80's or .90's. The phi coefficients computed for the items in the illustrative test are shown in Table 6, where they may be

TABLE 6

ITEM-TOTAL SCORE PHI COEFFICIENTS FOR ILLUSTRATIVE SIX-ITEM TEST

Item	ϕ_{it}	PPR_i	H_{it}
1	1.000	1.000	1.000
2	.375	.500	.714
3	.375	.500	.333
4	.167	.250	.619
5	.524	.667	.867
6	.524	.667	.889

compared with those computed by Jackson's PPR_i and Loevinger's H_{it}.

Though this method of computing a phi coefficient between a test item and the total score has the advantages of a known sampling distribution, absolute maximum and minimum values, and freedom from restrictive distribution assumptions, it does not furnish an index for the test as a whole. It is possible however to derive one by an averaging of the obtained phi coefficients. Such an approach is shown in Formula 20, which Cronbach says is analogous to Guttman's formula for reproducibility.

$$R=\frac{1}{K}\sum_{i=1}^{k} 1-2p_iq_i(1-\phi_{it}). \qquad [20]$$

Cronbach explains (**1**, p. 324):

> The correlation of any two-choice item with a total score on a test may be expressed as a phi coefficient, and this is common in conventional item analysis. Guttman dichotomizes the test scores at a cutting point selected by inspection of the data. We will get similar results if we dichotomize scores at

that point which cuts off the same proportion of cases as pass the item under study. [Our ϕ_{ii} will be less in some cases than it would be if determined by Guttman's inspection procedure.] Simple substitution in Guttman's definition . . . leads to [Formula 20 above] where the approximation is introduced by the difference in ways of dichotomizing. The actual R obtained by Guttman will be larger than that from [this formula].

In our example the value turns out to be .80 as compared with the reproducibility figure of .88.

This composite index for the entire test will have a maximum value of 1.00 and a chance value of

$$\frac{1}{K}\sum 1-2p_iq_i,$$

which approaches .50 as the average item difficulty approaches 50 per cent. The sampling distribution of this statistic, to our knowledge, is not known.

Discussion

This concludes the exposition of the major methods which have been put forward to give an index of the reproducibility of tests. Of those which yield indices for the test as a whole, several meet serious objections which have been leveled at Guttman's scalogram analysis. The techniques of Jackson, Loevinger, and Green are all objective, and result in measures which are not affected by the distribution of item difficulties. All have the same underlying rationale, but differ slightly in the way in which "errors" are counted. Loevinger's H_t is the most conservative of the three since all possible errors are counted; Jackson's PPR_t is the least conservative, and Green's I will usually fall between the two. The principal and not inconsiderable advantage of Green's technique is ease of computation, an important factor when the number of subjects and test items is large. Green's technique is the only one discussed that gives an estimate of significance for the reproducibility of the entire test.

Of the methods for computing the homogeneity of an item with the total test, the phi coefficient seems the most desirable because computation is easy and because the significance level of the obtained statistic can be determined exactly. Almost any of the commonly used item-analysis statistics—point biserial, biserial, or Flanagan correlation coefficient—may of course be interpreted as an index of item reproducibility, since in a reproducible test any person passing a given item will pass more other items than a person failing that item. They differ from the phi coefficient mainly in the number of assumptions they impose upon the data. Those employing conventional item-analysis statistics have been quite willing to assume an interval scale and a distribution function, usually normal, while those working within the framework of the concept of reproducibility have in general foresworn the unit of measurement and have thus confined themselves to distribution-free statistics.

All the reproducibility indices rest upon the same assumption that in a reproducible or homogeneous test, one can reproduce the entire response pattern of passes and fails, given the total number of items correct, and the item difficulties. All the methods employ the same data in the response matrix. All agree that in the response matrix of the perfectly reproducible test there will be no instances in which a subject passes an item more difficult than one he has failed. This is equivalent to saying either that all interitem covariances are maximal, or that the variance in total scores is maximal. Conversely, the test with lowest reproducibility will exhibit zero interitem covariances, and minimal variance in total scores.

Reproducibility and Factor Analysis

It is obvious that the phi coefficient method of determining the homogeneity of an item with total test is very similar to the procedure in classical test construction for "purifying" a test.

A common procedure for evaluating an item in conventional test construction is to compare the number of subjects passing the item among the 27 per cent of the sample making the highest total scores as opposed to the 27 per cent making the lowest total scores. A "good" item is one that discriminates between these highs and lows. Consequently, the items which would be chosen as producing the most reproducible scale in the phi procedure for obtaining reproducibility would also be selected as the most discriminating in conventional test statistics. This point is important when considering the relationship between reproducibility and factor analysis.

Several authors have been concerned with the question of the relationship between reproducibility and factor analysis. Loevinger (**14**) has stated that factor analysis and reproducibility are unrelated. Humphreys (**9**) appears to agree with Loevinger on this point and attacks reproducibility for not being as satisfactory a tool for research as factor analysis. He feels that reproducibility will lead to a confusing multiplicity of tests, while a factor analytic approach will not. Humphreys uses the hypothetical case of the problems involved in constructing a mechanical information test. The criterion of reproducibility, he fears, would require the construction of separate tests for the cross saw, the brace and bit, the pipe wrench, etc. On the other hand, all these tests would probably appear on a single common factor that would be orthogonal to other factors.

The writers disagree with both Loevinger and Humphreys, feeling that reproducibility and factor analysis are closely related. This relationship can be made obvious by consideration of the Wherry-Gaylord iterative analysis (**19**). This is a method for discovering homogeneous groupings of items in a test. It involves correlating each item with the total score. Items with the highest correlations are selected and the test rescored on the basis of these items. All the items are then correlated with the new total scores. This procedure is continued until a stable group of items is extracted. These items constitute a single factor. The remaining items can be rescored and additional factors extracted. The first factor removed would be the general factor. As can be seen, the phi method of obtaining reproducibility corresponds very closely to the Wherry-Gaylord extraction of the general factor. The principal differences are that the Wherry-Gaylord does not require that the finally selected items have a range of item difficulties, and does not cut total scores at the same ratio as the item-difficulty levels. Evidence reported by Wherry, Campbell, and Perloff (**20**) suggests that the Wherry-Gaylord general factor will correspond to the general factor obtainable in a Thurstone multiple factor analysis. The present writers found similar evidence in an analysis of a morale scale. After the morale scale had been subjected to a Thurstone multiple factor analysis, it was administered to a new group of subjects and subjected to a phi reproducibility scaling. The resulting scale was almost identical in item content with the Thurstone general factor.

While it appears to be true that a highly reproducible scale will tend to measure a single factor (since the phi analysis will isolate the general factor in the test items), not all single

factor tests will be highly reproducible scales. This is because a reproducible scale must have a range of difficulty levels if all persons are not to be forced into two categories: either all items passed or all items failed. The following example points up the reason this is true. If all items were at the 50-per-cent difficulty level, and if the test were perfectly reproducible, the 50 per cent of the subjects with the highest total scores would score correct on all items; the 50 per cent of the subjects with the lowest total scores would score incorrect on all items. This restriction is not necessary for all single-factor tests. Single-factor tests can have all items at the same difficulty level and still have a wide range of total scores due to the almost inevitable presence of error variance in the items. Reproducibility is impossible in such a case. Despite this lack of reproducibility, the single-factor test might be quite adequate since, if two persons score high on a single-factor test it is because they are high in the factor, and the differential patterns of their responses must be irrelevant for prediction since the differential patterns must be a result of error variance and do not represent stable patterns. If the differential patterns were differentially predictive, the test could not be a single-factor test. In those situations, therefore, where it is desirable to have all items at the same difficulty level, reproducibility is usually not a useful approach. The exception to this rule is the case in which a single discrimination is desired—e.g., pass vs. fail. In this case all items should have pass per cents which are proportional to the pass per cent desired for the whole test (**16**).

In many practical test-construction situations, where the logic of the situation is not incompatible with reproducibility, it appears to the writers that obtaining a general-factor test through phi reproducibility is simpler than through a Thurstone multiple-factor analysis. In addition to the relative ease of computation, the set of items so obtained should form not only a single-factor test, but also a reproducible scale.

Reproducibility and Reliability

It is obvious that the techniques for computing so-called reliability coefficients from a single test administration employ exactly the same data which have been used to compute the indices of reproducibility described above. Cronbach (**1**) has already pointed out the intimate relation of Guttman's reproducibility to the Kuder-Richardson Formula 20, which he has rechristened *alpha*. The key term in *alpha* is the ratio of two variances, $\sum pq/\sigma^2_x$. As Loevinger points out (**13**, p. 31) $\sum pq$ gives the raw score variance which would be obtained from a test whose items were completely independent, (σ^2_{het}); and σ^2_x is the obtained raw score variance. Loevinger's Formula 11 has these same quantities in it, plus a third representing the raw-score variance of a test whose items were perfectly correlated. It should be noted that the lower limit of *alpha* is always zero, but the upper limit is dependent upon the distribution of item difficulties. The obtained *alpha* for our illustrative test is .47, and the upper limit of *alpha* for this set of item difficulties is .88.

In order to make *alpha* independent of the distribution of item difficulties, Horst (**8**) has developed a formula which turns out to be identical with Loevinger's 11, except for a correction term composed of the ratio of the maximal to obtained score variance. Since this ratio has a lower limit of 1.00, figures obtained by Horst's method will necessarily be larger than Loevinger's except in the perfect case. The Horst formula for

the reliability coefficient corrected for dispersion of item difficulties is given below

$$r_{tt}=\frac{\sigma^2_x-\sum_{pq}}{\sigma^2_{max}-\sum_{pq}}\left(\frac{\sigma^2_{max}}{\sigma^2_x}\right). \quad [21]$$

The striking similarity of the Loevinger Formula 11 and the Horst Formula 21 cause one to suspect that the difference between single-trial reliability and homogeneity or reproducibility is more apparent than real.

The critical difference between the "reproducibility" and the "reliability" camps of test construction is seen most clearly in the ways they interpret their indices. When a test shows perfect reproducibility, it will also show perfect reliability by any of the formulas described so far. In order for this unlikely event to occur, several conditions must be met: all the items must be homogeneous in content, all subjects must be similarly constituted in the trait, attitude, or ability being tapped; and this trait, attitude, or ability must remain stable during the testing period. *Any* departure from these conditions will cause *any* of these measures to fall, and there is no way to tell on the basis of the response matrix alone what is amiss. An astute dropping of rows or columns from the matrix (subjects and/or items) will, of course, make things look better. In any event, a low figure indicates that considerable information will be lost in attempting to order subjects on a single linear continuum on the basis of their total scores. It is here that techniques such as Lazarsfeld's latent structure analysis (12) may be used to determine the minimal number of dimensions (classes) needed to account for the information contained in a response matrix. With this technique, a subject, instead of being given a total score, is assigned a probability of belonging in each of several classes. No unidimensionality is assumed, so there is no question of item or subject elimination to force unidimensionality, a procedure routinely employed by those addicted to Guttman scaling and classical test construction. When any such item-elimination procedure is used in test construction, a reliability or reproducibility figure computed on the final sample of items cannot be evaluated until the new version of the test has been administered to another sample of subjects. A low reproducibility figure is generally taken as an indication of item heterogeneity in a test, while a low reliability figure of the Kuder-Richardson variety is usually seen as an indication of the presence of considerable error variance. In the absence of other information, either interpretation is equally plausible, or suspect, since, as was pointed out above, the indices employ the same information from the response matrix.

Items and Subjects

There is no reason why the techniques of computing reproducibility or single trial reliability cannot be reversed to yield coefficients about the homogeneity of subjects, instead of test items. It is surprising that this has not been done more often, especially in the area of attitude measurement. Lack of reproducibility in a response matrix is just as likely to be due to heterogeneity in the population tested, as to heterogeneity in the test items. For most of the indices described above, computation of subject homogeneity would merely involve switching row and column marginals in the formulas. Such a technique would seem to be a promising one for the identification of deviants.

Why Reproducibility or Single-Trial Reliability?

Having come this far, it is high time we asked why a test with high

reproducibility or single-trial reliability is a good thing. Social scientists are all too prone to assume that it is, and to think no further about it. As Cronbach (1) has pointed out, reproducibility is in a sense a measure of the redundancy in a test. For many purposes, this is undesirable. Whenever test results are used to predict a dichotomous criterion such as hire–not hire, pass–fail, butcher –candlestickmaker, psychotic–normal—in short to classify subjects—it can be argued that the last thing in the world a test should have is high internal consistency. The real need is a set of items highly related to the criterion but not to each other. This is, of course, a restatement of the multiple-correlation approach to prediction. Ideally each item would represent a different pure factor. In such a situation, interest lies not in ordering subjects on some linear hypothetical trait, attitude, or ability continuum, but in an efficient dichotomization of the subjects or an ordering on the basis of the probability of membership in a class. To the extent that the test items are redundant, valuable testing time is wasted. It is a mistake to think such a test is "measuring" something, in the usual sense of that word. That a test can differentiate between neurotics and normals is no indication that "neuroticism" is a trait on which people can be ordered in some simple fashion. Much confusion in clinical literature is based on this fallacy. Unless the instrument exhibits high homogeneity-reproducibility–single-trial reliability, there is no reason to assume that the score on the test can yield an ordering of the subjects on some unidimensional continuum which can be given a label.

It is the person doing "basic" research who is apt to be more interested in ordering subjects on a unidimensional continuum. For him, the question of the internal consistency of his multiple-item test or questionnaire is of immediate concern. He may start with the unshakable conviction that the trait he has in mind *is* unidimensional, in which case he will engage in an often lengthy process of test construction, weeding out items until he achieves an instrument with internal consistency at a satisfactorily high level. This type of worker usually longs for an infinite population of items and subjects. When this longing is fulfilled, or even approximated, he can usually come up with a selection of items which, when administered to an appropriate population, will yield a response matrix of the desired internal consistency. He may even regard this achievement as support for his initial assumption about the unidimensionality of the trait, though the logic of such a conclusion is somewhat less than perfect, considering the amount of information thrown away in order to make things come out so neatly.

On the other hand, he may begin with a more modest aim: to find out, for a given set of items, the minimum number of parameters needed to account for the obtained responses of subjects to these items. If he finds that the response matrix shows high reproducibility or high single-trial reliability, he is apt to be pleased because life is so simple; but if he does not find his data so neatly arranged, he is likely to resign himself to fairly laborious procedures in order to find out the dimensionality of the data he has collected rather than to attribute any departure from unidimensionality to error variance.

The important point is that all the techniques mentioned here, whether they are regarded as indices of reproducibility, homogeneity, or single-trial reliability, are based upon the same raw data in the response ma-

trix; and all are more or less interchangeable with a little algebraic manipulation, though, as we have seen, they yield different numbers. How the number is interpreted depends not upon which one of these formulas is employed, since they are all basically equivalent, but upon what assumptions are made. One can assume that the items are homogeneous and that the subjects are similarly constituted in the trait being measured, in which case one uses the index as a measure of intraindividual trait stability. On the other hand, one can assume trait stability and subject homogeneity, in which case the index is said to reflect the homogeneity of the items. As was mentioned above, one may equally well assume trait stability and item homogeneity and employ the index as a measure of the homogeneity of the subjects. Any pair of assumptions appears to be about as plausible as any other. The important point is that from a single response matrix there is no way of telling what assumptions are reasonable. An obtained index, be it Jackson's PPR_t, Loevinger's H_t, Green's I, Cronbach's *alpha*, or Horst's r_{tt}, will be less than 1.00 when any or all of these conditions are not met. The plausibility of the assumptions can be ascertained only by recourse to further data, and the kind of data required will be different for testing each assumption. An estimate of intraindividual trait stability, for example, demands retesting the same subjects with the same items, but such retest data will be of little value in arriving at estimates of subject or item heterogeneity.

The one thing these indices of reproducibility or single-trial reliability will reflect without equivocation is the amount of information thrown away by representing the subject's performance on the test by a total score based on the number of items passed. They indicate, in other words, how adequately a unidimensional model fits the obtained data.

Proponents of homogeneity or reproducibility have been criticized because their criteria for a "good" test are unrealistically strict. It is true that perfect reproducibility will occur when, and only when: (*a*) the factors determining subjects' responses to the test do not change during the testing period, (*b*) the factors determining subjects' responses to the test are the same for all subjects, and (*c*) all the items in the test are identical in the factors determining the responses they elicit. It is also true that perfect single-trial reliability will be obtained only under the same circumstances. These are stringent conditions, and they are seldom, if ever, met. Human beings are just not that simple, but the fault is hardly Guttman's. There is nothing wrong in continuing to assume that many human abilities, attitudes, and traits are unidimensional continua, but we should be fully aware that this is at best a useful first approximation, and that an appreciable proportion of the information in our raw data will thereby be sacrificed on the altar of error variance.

REFERENCES

1. CRONBACH, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, **16**, 297–334.
2. FESTINGER, L. The treatment of qualitative data by "scale analysis." *Psychol. Bull.*, 1947, **44**, 146–161.
3. FORD, R. N. A rapid scoring procedure for scaling attitude questions. *Publ. Opin. Quart.*, 1950, **14**, 507–532.
4. GREEN, B. F. Attitude measurement. In G. Lindzey (Ed.), *Handbook of social psychology*. Cambridge: Addison-Wes-

ley, 1954.

5. GREEN, B. F. A method of scalogram analysis using summary statistics. *Psychometrika*, 1956, **21**, 79–88.
6. GUTTMAN, L. The Cornell technique for scale and intensity analysis. *Educ. psychol. Measmt*, 1947, **7**, 247–279.
7. GUTTMAN, L. The basis for scalogram analysis. In S. A. Stouffer et al., *Measurement and prediction*. Princeton: Princeton Univer. Press, 1950.
8. HORST, P. Correcting the Kuder-Richardson reliability for dispersion of item difficulties. *Psychol. Bull.*, 1953, **50**, 371–374.
9. HUMPHREYS, L. G. Test homogeneity and its measurement. *Amer. Psychologist*, 1949, **4**, 245. (Abstract)
10. JACKSON, J. M. A simple and more rigorous technique for scale analysis. In *A Manual of scale analysis*. Part II. Montreal: McGill Univer., 1949. (Mimeographed.)
11. KAHN, L. H., & BODINE, A. J. Guttman scale analysis by means of IBM equipment. *Educ. psychol. Measmt*, 1951, **11**, 298–314.
12. LAZARSFELD, P. F. The logic and mathematical foundation of latent structure analysis. In S. A. Stouffer et al., *Measurement and prediction*. Princeton: Princeton Univer. Press, 1950.
13. LOEVINGER, JANE. A systematic approach to the construction and evaluation of tests of ability. *Psychol. Monogr.*, 1947, **61**, No. 4 (Whole No. 285).
14. LOEVINGER, JANE. The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychol. Bull.*, 1948, **45**, 507–529.
15. LONG, J. A. Improved overlapping methods for determining the validities of test items. *J. exp. Educ.*, 1934, **2**, 264–268.
16. LORD, F. M. Some perspectives on "the attenuation paradox in test theory." *Psychol. Bull.*, 1955, **52**, 505–510.
17. MARDER, E. Linear segments: a technique for scalogram analysis. *Publ. Opin. Quart.*, 1952, **16**, 417–431.
18. NOLAND, E. W. Worker attitude and industrial absenteeism: a statistical appraisal. *Amer. sociol. Rev.*, 1945, **10**, 503–510.
19. WHERRY, J. J., & GAYLORD, R. H. The concept of test and item reliability in relation to factor pattern. *Psychometrika*, 1943, **8**, 247–269.
20. WHERRY, R. J., CAMPBELL, J. T., & PERLOFF, R. An empirical verification of the Wherry-Gaylord iterative factor analysis procedure. *Psychometrika*, 1951, **16**, 67–74.

Received May 12, 1956.

PSYCHOLOGICAL BULLETIN
Vol. 54, No. 2, 1957

AFFECTIVE PROCESSES IN PERCEPTION[1]

NOËL JENKIN[2]
The Training School at Vineland

Much imaginative and productive enterprise has for several years been expended in studying perception as it is related to a range of motivational functions. Activity in this field has almost, but not quite, acquired the character of a contemporary "school" (**20, 79**). It is not surprising, therefore, that this movement has become an object of critical attack as well as a focus for warm adherence.

Some current evaluations of its achievements and limitations are far from unanimous. M. D. Vernon (**123**) points out that in many experiments the long term schemata of the observer, by far the most important of the nonstimulus determinants of perception, are given no opportunity to function. She also notes that the relationship between "temporary need state" and perception has not been clearly established, and that even if the correlation exists, the results often may be attributable to a short-term cognitive set based on the actual conditions of the experiment. Such results, therefore, have less importance and generality than some would wish to suppose. Henle (**58**) has stated that the finding of a correlation between motivational conditions and performance on a cognitive task is only the first step toward solution of the problem. She finds thirteen possible ways in which needs and attitudes may influence cognitive processes, including perception, and calls for research aimed at specifying the particular manner by which, in a given experiment, a motivational state may act upon the percept. Murphy (**92**) has re-emphasized the importance of "autism," and of the learning process, in better understanding the nature of perceptual dynamics. Prentice (**109**) who has remained disenchanted by the broad "functionalistic" definition of perception, gestures invitingly toward the paradigm of the psychophysical experiment, and tells once more the story of the supposed failure of the new movement to explain *how* correlations of need and perception are mediated. That such correlations even exist is doubtful, he feels; they are "so hard to demonstrate."

Reflection of this nature may well leave many a reader with a sense of bewilderment concerning the field which is the target of such commentary. Since an ordered consideration of the data would probably yield clarification, it is the purpose of the present paper to organize a summary of the principal findings of recent years, and thus provide the most relevant materials for an assessment of the evidence. Two restrictions will be placed on the field to be covered. First, since the literature prior to 1949 has already been the subject of adequate review (**13, 25, 26, 29, 98, 99**), work done prior to this date will be omitted or dealt with by brief allusion. Second, in view of the magnitude of the area to be covered, work dealing with the perception of

[1] The author is indebted to Dr. D. W. MacKinnon, of the University of California, and to Dr. J. S. Bruner, of Harvard University, for reading critically the first draft of the present manuscript.

[2] This paper was prepared when the author was at Harvard University.

other persons, with perceptual typology and with "perceptual attitudes" or broad syndromes of related personality and cognitive functions, will be omitted. It is not thereby implied that this work is irrelevant or unimportant, but rather that a survey of the latter fields would best be undertaken as a separate task. The material reviewed will be grouped into four arbitrarily selected categories. The first consists of studies in which size judgment has constituted the dependent variable. The second is a group of investigations in which a physiological need (hunger or thirst) has been studied in relation to some perceptual activity. The next area to be considered will be the relation of positive values to perceptual behavior, followed by a review of work on the perception of noxious or threatening stimuli. A final section will deal briefly with the implications of the research previously reviewed, in relation to the problem of defining "perception."

Studies of Size Judgment

Little space is required to discuss the well-known Bruner and Goodman study (**25**), or the equally familiar counterfire by Carter and Schooler (**34**). The former found an enhancement of size which could be attributed to the desire for money, and "explained" in terms of perceptual accentuation. The latter found no such perceptual effect. Further work in different laboratories resulted in further inconsistencies. Bruner and Postman (**27**) found that tokens containing a "positive" symbol were judged larger than those containing a "negative" (unpleasant) one, and that both kinds of tokens were judged larger than the one containing a neutral design. Klein, Schlesinger, and Meister (**65**) in a similar type of experiment, failed to find any effect on size judgment from the affectively stimulating symbols inscribed on the stimulus objects. Solley and Lee (**116**) have recently reported further data, showing that when the stimulus figures are matched for closure, a valued object (in this case the object bearing a dollar sign) is judged significantly larger than neutral objects. No significant difference was found between judgments of the swastika and the neutral figure.

A different experimental design was used by Lambert, Solomon, and Watson (**69**). Judgments by young children of a disc, originally neutral in significance, showed an apparent enhancement and then a diminution in size, as the conditions of reinforcement and extinction were manipulated by the experimenters. Continuing this type of work, Lambert and Lambert (**70**) found similar results. Another study by Bruner and Postman (**24**) induced a different kind of affective state. While experiencing an electric shock, the *Ss* gave relatively accurate size judgments. Immediately after shock, however (the *Ss* now in a state of "tension-release"), significantly larger judgments were given.

A new approach to the question of value and the perception of size was made by Ashley, Harper, and Runyon (**4**). Fictional life histories of "poverty" and "wealth" were induced in hypnotized *Ss*, and significant results were obtained in the direction of those reported by Bruner and Goodman. An extensive investigation by Bruner and Rodrigues (**30**) returned to the area originally studied by the senior author (**25**), and attempted to resolve the differences between the Bruner and Goodman and the Carter and Schooler results. Differences in procedure between the

two studies had involved, first, shape of the variable patch of light which had served to obtain the measure, and, second, placement of the coins and discs which were to be matched. A further possible difference was that one group of *S*s may have adopted a set toward accuracy more strongly than the other group. The new experiment by Bruner and Rodrigues selected variables to represent these differences in a design intended to test the hypothesis that the value of objects will produce a constant error in the judgment of their size. It was found that coins were judged significantly larger than equivalent cardboard discs, and that there was no significant difference between the size estimates of coins and correspondingly-sized metal discs. This analysis represented the study of "absolute" accentuation, as in the earlier studies. A new method of computing the differential, termed "relative" accentuation, yielded the finding that as the value (and size) of coins is increased, the extent of overestimation increases significantly more markedly than is the case with metal or cardboard discs. Contrary to the prediction, "accuracy set" produced overestimation of both coins and discs, relative to performance of the group "value set." This finding refers to the condition where the objects to be judged were placed on the table before *S*. In contrast, when coins were held in the hand, greater relative accentuation was found for the value-set than for the accuracy-set condition. No significant difference was found between any of the three types of variable light patch used.

Lysak and Gilchrist (**81**) attempted to test the generality of some of the Bruner and Goodman findings, using a design which departs from the former experiment in two important respects. First, adult *S*s were used, and second, paper currency rather than coins was employed. A preliminary experiment established the fact that objects of the same size and shape as U. S. paper currency are progressively overestimated in size as a function of complexity of the impressed design. A second experiment found that there was no significant difference between the judgments of the size of a control rectangle and the judgments of one, five, and ten dollar bills, and that there was no trend toward increasing overestimation of the bills as their value increased. A group to whom the bills were "available" (i.e., who were told that they would be given the money if their judgments were accurate) made slightly larger judgments than the group to whom the bills were unavailable. Groups which judged the bills from memory, five minutes after being shown them, made smaller judgments than did the groups making matches with the bills in view. This was contrary to the prediction based on the Bruner and Goodman results. A third experiment, with a larger group of *S*s, confirmed these findings.

In most of the reported work on size estimation in relation to motivation, the former measure has been achieved by employing the method of mean error, or some adaptation of this classical technique. Dukes and Bevan (**42**) departed from this typical procedure and gave a recognition test after previously exposing *S*s to a gambling situation in which they won or lost a sum of money. To a significant degree, the greater the sum of money either won or lost, the greater was the size of the object chosen as matching the critical object which represented the extent of the

winnings or losses. A brief lapse of time took place between seeing the object and making the match. A further experiment by Dukes and Bevan (**41**) departs not only from the method of mean error but also from the otherwise uniform concern with the visual modality. With a modified method of constant stimulus differences, children made a series of judgments of weights. "Valued" weights were constructed by filling jars with candy, and "neutral" weights by filling jars with sand and sawdust. It was found that the valued objects were to a significant extent judged as heavier than the "neutral" objects.

Yet another technique was used by Beams (**8**) who selected child *S*s on the basis of their strong preference or dislike for certain kinds of food. A projected image of the food object was adjusted by the *S* until it appeared equal in size to the actual food object. The stimulus object and the matching projection were alternately monocularly observed. Larger judgments of the favored type of object were found in highly significant degree.

The problem of systematic individual differences in proneness to the size-enhancement effect has received relatively little attention. A preliminary report by Klein (**63**) gives emphasis to an earlier argument (**64**) about the importance of this area. The *S*s were preselected on the basis of performance on the Stroop color-word test, and thus were classified as "high-interference" and "low-interference" groups. Thirsty *S*s of the high-interference group, compared with satiated controls, underestimated the size of discs displaying thirst-related symbols. Overestimation was shown by the thirsty low-interference group. For combined thirst versus combined satiated groups the mean difference in performance was negligible.

Discussion

A survey of the findings and arguments, as they have in this area developed over the past nine years, leads first to the conviction that design and technique in such experiments has progressively improved. Perhaps some early faults have been replaced by other, and subtler, errors. It is noteworthy, however, that the weight of evidence favors the proposition that value and need are determinants of size judgment. Not only do the experiments with positive findings outnumber the negative ones; it can also be said that most of the recent and best-controlled experiments are among those with a positive outcome.

In an early critique, Pastore (**97**) was able to argue that the relative overestimation in the Bruner and Goodman experiment may be a function of the size of the coin rather than of its value. This cannot be said of the more recent Bruner and Rodrigues experiment, with their new method of calculating relative accentuation. Nor can it be said of various other findings, which employed stimuli other than coins.

Reviewing some of the earlier work, with its puzzling and vexatious inconsistencies, Bruner and Postman (**28**) were prompted to ask: "What kinds of constraints operate in the stimulus field which enhance or inhibit the operation of directive factors? And what kinds of instructional or motivational constraints operate?" Implied here is the suggestion that there were unrecognized variables present in the earlier experiments which were uncontrolled and which led to the inconsistencies between

different studies. The Bruner and Rodrigues experiment achieved some success in defining and holding constant a greater range of variables. Under these conditions it was still possible to conclude that the value of objects affects their phenomenal appearance. Lysak and Gilchrist (**81**) however, holding constant the size of the object to be matched, obtained clearly negative results, and in attempting to account for inconsistencies in this field of investigation, they propose a developmental hypothesis. The progressive equivocality of experimental findings as the age of *S*s increases suggests that progress toward maturity brings an increasing ability to evaluate the physical environment. Nevertheless, several of the findings reviewed above have shown the "accentuation" effect with adult *S*s, and hence the developmental hypothesis, at least in its simple form, does not account for all of the published results.

Gilchrist and Nesberg (**52**) question the appropriateness of procedures in which the standard and the variable stimuli have differed in dimensions other than that in which the *S*s were required to make their matches. The experiment of Beams, cited above, attempts to avoid this difficulty, and in large measure probably succeeds. A new difficulty is consequently introduced, however. With his method, the stimuli are no longer simultaneously present in the *S*'s binocular field of vision. The brief transition from stimulus object to comparison object involves an interval of time which, though it is small, renders possible the objection that it is a memorial rather than a perceptual phenomenon which is being studied. Whether or not this objection is valid, the experiment seems certainly to confirm further the view that accentuation effects as a function of need and value are shown most strongly when some degree of equivocality is present in the stimulus situation. It is possible, as Bruner points out (**21**), that optimal viewing conditions may eliminate the effect altogether.

Finally, brief reference should be made to the oft-raised question of *why* valued or needed objects or those associated with pleasant affect should in some dimension be perceived as greater than neutral objects. Bruner and Rodrigues (**30**) suggest that the effect is due to the frequent pairing in the environment of value and size. An alternative suggestion is made by Dukes and Bevan (**41**). Their previous work (**10**, **41**) had shown that accentuation effects, in the case of valued objects, are coupled with decreased variability of response. This led them to adopt an analogy from the field of electronics. Motivational factors, such as needs and values, may serve to "tune" the organism to respond with high selectivity and amplification (accentuation), when it is in the presence of valued stimulus objects: When the receiving system encounters less valued objects, the perceiver responds to a wider range of stimulation, but at the same time sacrifices the degree of amplification which occurs with sharp selectivity.

Physiological Need and Perception

The factor of ambiguity in the experimental situation has generally been minimized in the studies discussed above. In this respect, they are to be distinguished from practically all other experiments in the field under review. The contrasting type of work has sought to isolate perception-motivation relationships by

means which include either the reduction of stimulation to some critical value or the broadening of a range of response probabilities, or both. Typical of this approach is a group of experiments in which either the hunger or thirst drive has been manipulated as an independent variable. Influenced by the work of Sanford (**114**), several of these studies used what has been styled as a "projective" technique (**129**).

Sanford reported that the number of food responses made by hungry *Ss* in several different situations increases in a negatively accelerated manner. This finding provides a context for the consideration of studies more directly and explicitly dealing with the perceptual process. The same is true of several other studies, including an experiment by McClelland and Atkinson (**84**). These investigators also found that responses of a food-related character increased in number as the period of food deprivation lengthened. The "projective" technique in this experiment achieved the ultimate in stimulus ambiguity; in some of the trials a completely blank screen was used and the *Ss* were led to believe that faint visual cues were present. In a second experiment (**5**), TAT stories were evoked under different degrees of hunger motivation. Although a decrease occurred in the frequency of references to eating as the interval of deprivation lengthened, certain other trends were noted. Plots tended to involve the desire for food or activities designed to remove the obstacle in the way of hunger satisfaction.

From these experiments, it is possible to draw a conclusion that food-deprivation has a marked effect upon cognitive processes, possibly including perception. This generalization must be qualified, however, by the negative outcome of the experiment of Brozek, Guetzkow, and Baldwin (**18**). The hunger need was aroused in this experiment more drastically than in any other work reported. A state of semistarvation was maintained in the *Ss* for 24 weeks. Despite very clear "clinical" evidence that thoughts of food came to preoccupy, and indeed even to obsess the minds of the *Ss*, little or no relationship was found between the independent variable and the projective-test measures employed, including the Rorschach and the Rosenzweig Picture-Frustration Tests. One result, separately considered, was of statistical significance. The experimental group made a higher mean percentage of idiosyncratic responses per word to eight food words from the Kent-Rosanoff list.

Also within this context of work closely related to hunger and perception, is the experiment of Postman and Crutchfield (**105**). They required *Ss* to supply missing letters in words which offered opportunity for completion as food or nonfood words. The degree of hunger in the *Ss*, the degree of ambiguity of the words, and the degree of selective set for responding with food words were systematically varied. For the most hungry of the groups, the increase in food responses as a function of degree of set was positively accelerated. For the nonhungry group the increase was negatively accelerated. As deprivation was prolonged, there was a decrease in the number of food responses to the least ambiguous stimulus words and an increase in food responses to the most ambiguous words. Michaux (**90**), using a technique similar to that of Postman and Crutchfield, confirmed his prediction that a group of persons with no history of mental disorder would show when

hungry an increase of "apperceptive emphasis on food," while a group of schizophrenic patients would, when hungry, fail to manifest this trend. The finding is interpreted in terms of a defect of "psychological homeostasis."

Work showing or attempting to show a relationship between hunger and perception, as distinct from related cognitive processes, appears to have begun with the experiment of Levine, Chein, and Murphy (**76**). Ambiguous drawings were presented to hungry *S*s, who yielded more food-related responses than were given by a control group. The food-related responses to achromatic drawings increased in number after three hours of food deprivation, and increased still further after six hours, but decreased after nine hours of deprivation. For chromatic pictures, the increase occurred at three hours and the decrease at six hours.

Gilchrist and Nesberg (**52**), in attempting to secure an unequivocal answer to questions about the relationship of need and perception, abandoned the "projective" method, and embarked on a strictly controlled series of experiments in which hunger and thirst were the independent variables. Their *S*s for 15 seconds observed the projected images of food and drink objects immediately after a meal, and again at 6 hours and at 20 hours after eating. The light was switched off for 10 seconds and the *S* was then required to adjust the brightness of the image to the degree of illumination previously seen. Hungry *S*s made significantly brighter matches than the controls, and this effect increased as a function of the time of deprivation. A second experiment, using thirst as the independent variable, found similar results. The illuminance matches of the experimental group rose steeply with the period of deprivation, then showed some negative acceleration. Since the control group showed no significant change, the interaction reached a high order of significance. Two additional experiments confirmed these results while controlling for the possible influence of stimulus factors irrelevant to the thirst need. From the four experiments, it was concluded that support is given to the proposition: "Increasing need gives rise to an increasingly positive time error in the illuminance matches of objects relevant to that need."

A report by Lazarus, Yousem, and Arenberg (**75**) criticizes some earlier work in this field for failing to define perception in terms of the identification of objective stimuli. Two kinds of interrelated perceptions are proposed, one which seems to be more oriented toward imagination, association or projection, and one which is more stimulus oriented. In order to study "perceptual behavior in the strictest sense," two experiments were conducted in which unequivocal pictures of food and nonfood objects were shown for one-fifth of a second at progressively increasing degrees of illumination. The *S*s were free to guess at the identity of the objects. Thresholds for food recognition, relative to thresholds for nonfood recognition, decreased at 2 hours and at approximately 4 hours after deprivation but increased sharply at 6 hours. A replication of the experiment gave similar results. A further experiment (**75**) was identical in design, except that a forced-choice technique was used with a limited range of alternatives always before the *S*. In this situation, no significant relationship was found between hunger and the perceptual recognition score. This is interpreted as evidence "that need in

perception relationships depends on a wide range of response opportunities." In both experiments, a study of prerecognition guesses yielded no support for the hypothesis that response availability would account for the relationship between hunger and perceptual recognition.

The three perceptual studies discussed above have uniformly employed stimuli directly representing food and drink objects. Their positive outcome might prompt the question as to whether the effect of need would still manifest if, instead of direct representation, linguistic symbols for food and drink were used. A positive answer is offered by Wispé and Drambarean (**129**). Two groups of *Ss* deprived of food and water were compared with a nondeprived control group as to their respective recognition thresholds for "neutral" words and words related to hunger and thirst. In order to study the effect of different frequencies of word usage, two lists of need-related words were separately given prior standardization and matched for frequency, one a list of common words and the other a list of uncommon words. Lowered thresholds for the deprived groups were found to a significant degree, for both common and uncommon need-related words. The relationship was not linear, as shown by the fact that the thresholds after a 24-hour interval were not lower than those after a 10-hour interval. A study of the prerecognition responses found that words relating to food objects and to acts instrumental to need-satisfaction increased in frequency at 10 hours and decreased at the 24-hour interval.

A recent experiment by Taylor (**121**) along very similar lines found results which conflict with those outlined above. Degree of need and degree of "set" in the *Ss* were both manipulated, half of the satiated and half of the nine-hour deprived groups both being given instructions which led them to expect words referring to food or beverages. Those *Ss* given "set" instructions showed lower thresholds for need-related words than did the nonset groups, but there was no significant difference between the thresholds of the deprived and satiated groups, even for the subjects not given "set" instructions. A replication of the experiment with a different ordering of the stimuli showed the same negative results.

Discussion

Of the six reports upon hunger in relation to a broadly defined cognitive area, only one (**18**) was essentially negative, and this was the one which used stimuli largely irrelevant to the need-state studied. It seems from the empirical findings that when appropriate stimulus objects are used, the presence of the hunger need facilitates categorization in a manner consistent with that need. The same findings compel us to note that "appropriate stimulus objects" constitute a large and varied class, ranging from incomplete words to a blank screen. Further, we must incongruously exclude from this class the Rorschach blots and the Rosenzweig Picture-Frustration Test (cf. Brozek, et al., **18**). Some measure of congruity of the stimulus object with the need is evidently one factor which produces the effect. On the other hand, a very high degree of ambiguity (e.g., a blank screen plus suggestions that cues are present) is also a favorable condition for the occurrence of need-related responses.

The crucial question for the present discussion must ask if the relationship between need-state and such processes as imagery, association, and

problem solving extends also to the perceptual process itself. From the results of four out of the five perceptual studies reviewed, an affirmative answer is indicated. The earliest of these (76), has been several times criticized on methodological grounds (2, 75, 97), but even with the exclusion of this instance, the weight of evidence seems to favor the hypothesis. Of the remaining studies, two (121, 129) used stimuli which were only indirectly representative of the goal objects (i.e., words rather than pictures) and their results are in conflict. Two studies remain (52, 75) which offer clear evidence for a relationship between need-state and perception. The principal independent variables in both groups of experiments were similar, yet they used entirely different measures of performance—recognition thresholds for pictures versus illuminance matches of pictures. Considered in the context of research in the same general area, including the work on size estimation, these results are therefore convincing.

As in the previous section, it is necessary to ask why need variables should apparently function as determinants of perception. A feature of possible significance in the solution of this problem lies in the shape of the function graphically plotted between degree of deprivation and the dependent variable. Despite widely differing kinds of measures, all experiments having a positive outcome in this type of perceptual and perceptual-cognitive research have shown similar features. As hunger has increased, the plot has shown either a U curve or one of negative acceleration. This effect is at minimum, though still discernible, in the Gilchrist and Nesberg (52) experiments and at a maximum in the Levine et al. (76) and Lazarus et al. (75) studies. In the Postman and Crutchfield study (105), it is true only for the "low-probability" list, i.e., words less likely to evoke "food" solutions.

Relevant to this common finding is the hypothesis of McClelland (83) who proposes, in terms of personality theory, an explanation of the curvilinear relationship. Motivation is presumed to have different effects at different intensity levels. When it is weak, in the "wish-fulfillment" stage, goal images occur. As it increases, a "push toward reality" is experienced, in which *deprivation* imagery tends to replace goal imagery. Still further increase in motivation brings orientation toward relief from anxiety rather than toward attainment of the original goal, and in this stage occurs "a kind of defensive goal imagery which is very different in its function from the goal imagery obtained with weak motivation."

An alternative explanation to that of McClelland is offered by Lazarus et al. (75), who follow a similar view proposed earlier by Sanford (114). Since sensitivity to the food objects has been shown to increase after about 3 or 4 hours of deprivation and then to level off or decrease, it is suggested that the perceptual curve follows the cyclical food habit. Support for this hypothesis would require evidence of an increase in sensitivity *after* the initial rise and fall. Experimenters (e.g., 52, 129) who have used longer periods of deprivation than did Lazarus et al. have not as yet reported any such recurrent rise in the function measured.

Wispé (128) suggests that the increase in need-related associations occurs initially as a result of a "food

habit" rather than real tissue need. It is implied that the influence of the "food habit" declines as the period of deprivation is protracted, leaving, however, the state of physiological need to operate as the determinant of a higher than normal level of responding with food associations.

Much of the interest of the experiments on hunger and thirst in relation to perception lies in the possibility that changes in body chemistry are closely related to perceptual experience (7). Test of a two-component theory, such as that proposed above, would require experimental separation of the habit or appetite variable from the physiological need. The work of Beams (8), cited in the previous section, suggests a possible method for inclusion of the former variable in a multifactor design intended to analyze the complex functions which are evidently operating.

Until such work is conducted, the experimental data dealing with reduced thresholds for need-related objects and symbols are best interpreted, it seems, in a way which is probably inadequate and which may be only partially true. This postulates a learned association between need-state and need-related responses, with the consequence that the latter have an increased probability and facility of occurrence under appropriate drive conditions and in an appropriate stimulus situation. Such an explanation does not readily explain the Gilchrist and Nesberg (52) finding of a positive time error for illuminance matches of need-relevant objects. This type of perceptual "accentuation" is more akin to the kind of data considered under the heading "studies of size judgment" and might best be understood in the context of those experiments.

Positive Values and Perception

The stimulus for a great deal of subsequent research was a paper by Postman, Bruner, and McGinnies (101), published in 1948. Entitled "Personal Values as Selective Factors in Perception," this reported an investigation prompted by the question: what does the individual contribute to perceptual selection over and above a healthy pair of eyes and the appropriate response mechanisms? Measures on the Allport-Vernon Study of Values were compared with *S*s' recognition thresholds for tachistoscopically exposed words which represented each of the value areas. The results were discussed in terms of concepts developed by Bruner and Postman (23) in a previous study—selective sensitization and perceptual defense. To these was added a new concept, that of value resonance, based on a study of the *S*s' presolution hypotheses.

A special feature of the Postman, Bruner, and McGinnies experiment lay in the fact that the main independent variable was not a generalized motivational state, assumed to be uniform for all *S*s. Instead, a certain range of individual differences (the Spranger values) was selected and the measures on the dependent variable (recognition of value words) were treated with respect to the ordering of each individual *S*'s value profile. This enterprising step opened up a new path for research, in which the focus upon perception was directed through the lens of personality, rather than through the broadly defined motivational state.

This innovation was adopted by subsequent investigators, among the first of whom were McGinnies and Bowles (87), who used as stimuli portraits rather than words. A value was

attached to each face by telling the *S* in each case that: "This is a scientist (or artist, minister, etc.)." The occupations thus denoted represented each of the Spranger values. The *S*'s score was the number of exposures necessary for him correctly to identify each of the faces. Correlation coefficients were computed for each individual *S* between his scores for identifying each occupational representative and his scores on the Allport-Vernon Study of Values. Negative correlations were obtained for 15 of the *S*s; the remaining 9 were positive. A further analysis showed a close rank-order agreement between value rank and the total number of correct identifications on the first recognition trial. It was concluded that when the experimental design does not offer opportunity for reduced thresholds for valued stimuli, "selective sensitization" manifests itself in greater ease of fixating visual symbols of preferred values. Although the majority of *S*s learned more easily to recognize valued symbols, it was felt to be of some significance that a smaller group found this task more difficult. A parallel was seen in the similar findings of previous writers (**23**, **101**) who interpreted special sensitivity to less valued symbols as "selective vigilance"—the antithesis of "perceptual defense."

Vanderplas and Blake (**122**), seeking to extend the general validity of the concept of perceptual sensitization, designed an experiment which varied from the Postman, Bruner, and McGinnies study in that the auditory rather than the visual modality was employed. The authors were able to demonstrate that auditory perceptual sensitization operates differentially to raise or lower recognition thresholds in a manner consonant with individual values as defined and measured by an independent instrument. A small minority of the *S*s again showed the opposite trend, i.e., "vigilance" for the words representing low-ranking values.

The conclusion reached by Postman, Bruner, and McGinnies was questioned by Solomon and Howes (**61**, **117**) who reinterpreted their results on the basis of a word-frequency hypothesis. It was assumed first that "high valuation of a given area of interest is associated with a positive deviation from the mean frequencies with which words in that area occur in general usage" (**117**). Second, it was proposed that the Allport-Vernon test itself can be considered a measure of the frequency with which the *S* uses certain words, and that it is unnecessary to postulate entities such as "values" in order to account for an *S*'s profile. The experiment reported by Solomon and Howes employed words of two categories, relatively frequent and relatively infrequent, according to the Thorndike-Lorge count. For each value rank, the data showed lower recognition thresholds for the frequent words than for the infrequent words. Between value rank on the Study of Values and recognition thresholds, a statistically nonsignificant trend was found in the direction reported by the former writers (**101**). These results were regarded as evidence consistent with the propositions cited above.

Postman and Schneider (**104**), after communication with Solomon and Howes, published simultaneously with the latter authors, a further report on the same type of data. Their stimulus words were also grouped into categories of relatively high and relatively low frequencies of occurrence. Two differences from the

Solomon and Howes lists may be distinguished. First, the words were classified as falling into the value areas by means of a consensus of three judges "thoroughly familiar with the test." Second, the "unfamiliar" words chosen were considerably *less* unfamiliar than those of Solomon and Howes. Hence, while the latter experimenters chose words such as "percipience," "erudition," "uncoerced" and "vignette," Postman and Schneider chose for the unfamiliar list examples like "conception," "logic," "dominant," and "literature." In general, the results of the experiment confirmed the finding that high-frequency words are recognized more easily than low-frequency words and that for high-frequency words there is no systematic relationship between value rank on the Allport-Vernon test and recognition thresholds. For low-frequency words, however, the relationship with value rank was present and statistically significant, the direction of the trend being again in the direction found by Postman, Bruner, and McGinnies.

What seems to be an effective rebuttal of the Solomon and Howes argument was presented by Adams and Brown (**1**) and Brown and Adams (**17**). In the latter paper, an experiment is reported which tests the Solomon and Howes hypothesis that results from the Allport-Vernon test are accountable for in terms of word frequency. The test was revised in such a way that the alternatives in all areas except one were expressed in synonyms with a very low frequency of usage. Six forms were constructed, each one favoring frequency-wise a different value area. On the Solomon and Howes hypothesis, the *S* should choose the high-frequency words and thus achieve a high rank for the corresponding "value." Six groups of *S*s answered one form of the new test, and also the Allport-Vernon-Lindzey test. It was found that there were no consistent changes in scores on the value area emphasized for frequency, relative to the five other groups in the same value area of the new test. Further, correlations of the new test with the Allport-Vernon-Lindzey scale remained significantly positive, notwithstanding the changes in frequency in the six versions of the former. It is concluded that the results disprove the Solomon and Howes hypothesis and are consistent with "the postulation of a central cognitive affective construct which may be called 'value area.' "

Haigh and Fiske (**56**) with an improved statistical procedure, have also repeated the Postman, Bruner, and McGinnies study (**101**). They style the use of the Allport-Vernon test by the previous authors an "indirect" measure of value preference, and supplement it in their own work by a "direct" measure, which consists of a ranking by each *S* of the 36 words shown to him tachistoscopically. The rank order was obtained within four weeks after conducting the perceptual experiment. The results obtained by the previous experimenters were corroborated by use of the "indirect" method. Use of the new "direct" method also gave confirmation but with a higher level of statistical significance. It is concluded that positive values tend to be associated with shorter recognition times.

Two criticisms of the Postman, Bruner, and McGinnies position were made by Mausner and Siegel (**89**). One was the word-familiarity argument, discussed earlier in this section, and the second was based on the view that the Allport-Vernon instrument is an inadequate test of values.

Seeking a situation in which the factor of familiarity was controlled and wherein value was varied in a simple manner, these authors designed an experiment in which adolescent stamp collectors were induced to learn the monetary "values" (i.e., alleged worth according to a firm of stamp merchants) of the various members of a set of stamps. No significant relationship was found between recognition thresholds for the stamps and their respective "values." The results were interpreted as evidence failing to support the Postman, Bruner, and McGinnies hypothesis. It should be noted, however, that the term "value" was in this experiment employed in a sense different from the Allport-Vernon usage, and that learning scores were not reported. Inspection of the data shows a distinct, though statistically nonsignificant trend in the direction of lower thresholds for higher-valued stamps. Use of additional controls and a more sensitive statistical test might well have resulted in the hypothesis' being supported.

A method different from that of the studies discussed above was devised by McClelland and Liberman (**85**). On the basis of combined TAT and performance-task measures of n achievement, *S*s were grouped into high, middle, and low categories. Three months after the personality measures had been secured, the *S*s were tachistoscopically presented with verbal material. Relative to the threshold for neutral words, the high n-achievement group perceived achievement-related words significantly more easily than did the middle and low groups. Security-related words were perceived significantly more easily by the middle and high groups, as compared with the performance of the group low on n achievement. The authors comment that the average familiarity ranking of the words employed as stimuli was sufficiently close to make highly implausible any interpretation of the results in terms of differential familiarity. A parallel type of study by Lindner (**77**) showed that "sensitization" for ambiguous, sexually-suggestive picture material was greater for a group of sexual offenders than for a control group of nonsexual offenders. Though the responses were not of a socially approved character, they presumably reflected the positive "values" of this special group.

Results difficult to interpret in the present context are reported by Gilchrist, Ludeman, and Lysak (**53**). Groups of students representing the extremes of a distribution on an anti-Semitism scale were used as *S*s. Positively valued, negatively valued, and neutral words were used as stimuli, each of which appeared on two slides, one containing the word "ink" above and below the stimulus word, and the other the word "Jew." A context was thus provided, but the *S* was asked to report only the stimulus word. It was found that both positive and negative values lowered word-recognition thresholds in comparison with neutral value, and also that emotionally loaded context has the effect of raising the thresholds for both positively and negatively valued words, while lowering the thresholds for neutral words. These results cannot be explained in terms of the word-frequency hypothesis, since the positively and negatively valued words were matched for frequency and the neutral words were actually chosen from a higher frequency category. The authors point out that these results pose problems both for the concept of "response suppression" and that of "perceptual defense." The

same is true of the concept of selective sensitization, unless this be broadened to refer to negatively valued, as well as positively valued stimuli.

Discussion

An overview of the work on selective sensitization leads compellingly to the conclusion that the concept is still a valid and useful one, and that the phenomena described by this term are not artifacts of word frequency or any other spurious variable yet distinguished. What remain to be clarified are the conditions under which sensitization occurs. A start has been made in this direction (**104, 117**), enabling us now to say with some assurance that preferred personal value and *moderate* unfamiliarity of relevant words are conditions which produce the effect, when these words are exposed in isolation. One experiment (**53**) has indicated that when the stimuli are simultaneously presented with a context of other words, sensitization occurs for words of negative as well as positive value. This enigmatic result is inconsistent with the bulk of evidence. It resists any plausible interpretation and clearly calls for further study.

One possible direction for future research is pointed out by the instances (**6, 101, 122**) in which some *Ss* have functioned in a manner opposite to that predicted. The treatment of such data in terms of "selective vigilance" by the same writers who proposed the concept of "perceptual defense" has been criticized as an inconsistency (**2, 62**). Postman (**100**) has replied, showing that these concepts were not invoked as explanatory principles and that the situation contains no more inconsistency than the parallel one in learning theory, where antithetical principles of facilitation and inhibition are found useful. If it is a fact that some individuals consistently tend to show "sensitization" for positively valued and others tend to show "vigilance" for negatively valued stimuli, and if means can be devised to predict in advance which individuals will behave in these respective ways, the implications will be important. A possible personality dimension is herein suggested, which may provide a basis for more effective control in value and perception research. Though such a control is hypothetical at present, it can be seen that if it should be developed and applied, there may then be avoided the kind of ambiguity inherent in the data and conclusions of Gilchrist, Ludeman, and Lysak (**53**). On the other hand, the so-called "vigilance" may be a function of one or several unrecognized variables present in the experimental situation. In either case, the facts call for further elucidation.

A surprising feature of the area reviewed above is that with few exceptions (**53, 77, 85**), all investigators using personality variables have chosen one instrument—the Allport-Vernon Study of Values. The same logic which prompted the Postman, Bruner, and McGinnies experiment could equally well have led to the selection of dominance-submission, extraversion-introversion, egocentricity-altruism or a host of other attributes of personality. In short, what McClelland and Liberman have attempted with n achievement remains to be done with other variables also. One probable reason why the field has not been further explored is the feeling that such correlates of perception, in and of themselves, are of little value. As Bruner (**20**) remarks, such data serve not to explain perception but to indicate problems. One

such problem has been raised by the correlation between the Study of Values and certain recognition thresholds. Essentially, it is the question of a mediating mechanism between the trait (or "value") and the percept. What is now needed is close study of a variety of conditions under which the correlation appears and fails to appear. The Study of Values is not necessarily the most sensitive or reliable instrument for such a search. Thus there is room for a great deal more exploratory work in the area considered. When the most satisfactory personality measure has been determined, the way will be clear for an intensive study of the factors responsible for the correlation.

Reactions to Noxious Stimuli

The area of personal values, just considered, has been somewhat arbitrarily separated from the much more extensive line of work which has been concerned primarily with the perception of noxious, "inimical," threatening, or "taboo" material. The finding of raised thresholds in relation to such stimuli, as compared with thresholds for "neutral" objects, was a phenomenon to which Bruner and Postman in 1947 attached the term "perceptual defense" (**23**). This and subsequent work by these writers (**19**, **101**), and also by McGinnies (**86**), established this concept in the literature, aroused unusual interest and led to a surprising amount of debate and dispute, some of which tended to be acrimonious.

The concept of perceptual defense was not proposed by the original authors as an "explanatory" principle, and it was made clear that work was needed to uncover the mediating mechanisms involved (**28**, **101**). Notwithstanding this caution in presenting the idea, some psychologists reacted with concern to the possibility that an "homunculus" process was implied (e.g., Howie, **62**). It is clear that this assumption was not justified. A more serious objection, however, was made by Solomon and Howes (**117**), who claimed that two simple processes account for the data of these experiments. The first is the frequency hypothesis discussed in the preceding section, and the second is the view that responses to taboo words are not delayed in perceptual recognition but merely delayed in verbal report. Other writers (e.g., Whittaker, Gilchrist, and Fischer, **126**) have added weight to one or both of these arguments and have also proposed explanations in terms of selectively reporting sets.

Postman, Bronson, and Gropper (**107**) posed the question in a general way: Can perceptual defense be reduced to the operation of determinants which are not specifically emotional? An answer was sought by designing an experiment in which taboo and neutral words were matched for frequency, and four different sets of instructions were used in order to manipulate the *S*s' readiness to report taboo words. A further attempt to vary the factor of selective verbal report was made by systematically varying the sex of *E* and the sex of *S*. Under all conditions, the thresholds for taboo words were found to be lower than for the neutral control words. This was thought to be due to a systematic underestimation of the familiarity of the taboo words. Relative thresholds for the two types of words varied significantly with the nature of the instructions, in the direction of the naive group having higher thresholds than any of the groups forewarned to expect taboo words. On the basis of this finding

and from a review of former studies, it was concluded that "perceptual defense has, at best, the status of an unconfirmed hypothesis." A similar experiment by Lacy, Lewinger, and Adamson (**68**) showed that the factor of expectation acted to reduce recognition thresholds more rapidly for taboo words than for neutral words, and also that an habituation effect rapidly leveled the threshold differences which occurred early in the series between the two classes of words.

In an experiment similar to that of Postman, Bronson, and Gropper, it was shown by Freeman (**50**) that when *Ss* are set to look for and report taboo words, their thresholds for these words are not higher than for neutral words. Continuing this study, Freeman (**51**) measured recognition thresholds of separate groups of male and female *Ss*. Neutral and taboo words were again used, and half of the *Ss* in each sex group were informed that the stimulus list would contain some taboo words. Typically, raised thresholds for taboo words were found in the uninformed group and little mean difference between the thresholds for the two classes of words in the informed group. Sex of the *Ss* played a significant role, informed females showing less reduction of the taboo word threshold than did informed males. A further experiment (**51**) was identical in design but used for the experimental group "ego involving" instructions which led *Ss* to believe that the perceptual task was related to academic success and aptitude. Ego involvement had the effect of reducing thresholds for all words (neutral as well as taboo) and this effect was much more pronounced for females than for males. In place of the "perceptual defense" interpretation, Freeman (**50**) proposes that raised thresholds occur "as a function of the dominance of alternative sets which do ***not*** predispose *S* toward the perception of taboo material."

Several experiments have been reported which are claimed to demonstrate perceptual defense while controlling for the factors revealed in the previous group of studies. Representative of such work is a series of studies by Cowen and Beier. Their first experiment (**38**) required a group of *Ss* to examine a booklet in which decreasingly blurred versions of a single word appeared as the pages were turned. Several such booklets provided stimulus material consisting of threat and nonthreat words. A second group followed the same procedure, except that it was "alerted" to a threat experience by prior exposure to the words it would subsequently decipher. More time and trials were required to report the threat words than the nonthreat words under both alerted and nonalerted conditions, although the difference was significantly greater for the latter condition. Both group and individual variability in responding to threat words increased under alerted conditions. A subsequent experiment (**9**) using the same technique confirmed at a higher level of significance the finding that *Ss* "though alerted to possible threat, nevertheless respond less accurately and less promptly to threat words than to neutral ones." A third experiment (**39**) again confirmed this finding while controlling specifically for the variable of word frequency and also for social setting, insofar as the latter involved sex roles. The writers consider that an explanation of their results in terms of conscious inhibition would be inadequate, and that word frequency must be excluded from any interpretation of the findings.

Another approach to the problem

was made by Newton (**95**) who equated two lists of words for frequency, and found that under conditions of tachistoscopic exposure, significantly fewer errors of recognition were made of the pleasant than the unpleasant words. Since the latter were not of the "taboo" variety, it was felt that the probability of response-suppression was reduced to a minimum. Wiener (**127**) controlled for frequency by using the identical stimulus words as "threat" and "nonthreat" stimuli. This was achieved by embedding the words in contexts which supplied different meanings for the different groups. Selective set was controlled by the use of "neutral" stimulus words in addition to the critical words in a neutral context. The "threat" group required significantly fewer trials than the "neutral" group to report the critical words correctly. The experimenter claims that while the direction of the difference is opposite to that shown by much other work on perceptual defense, this evidence is clearly in favor of motivation as a determinant of perception. A subsequent experiment (**33**) clarified the former finding by distinguishing two groups of *Ss* on the basis of clinical criteria, and predicting that they would show either sensitization or defense in the perceptual situation. Those classified as "sex sensitizers" perceived sexual words with significantly fewer trials than did those classified as "sex repressers." A similar finding was reported for those classified in regard to hostility. A third dichotomized group, distinguished in terms of consciousness of personal adequacy, differed in the predicted direction without reaching the criterion for statistical significance.

Eriksen (**46**) has also defended the notion of perceptual defense, though he justly criticizes the methodological errors of the "dirty word" procedure. His own technique has been to study the phenomenon in relation to individual differences as distinguished by clinical methods. In this he supports the argument of Lazarus (**72**) who points out in reply to Postman, Bronson, and Gropper that repression is not the only mechanism of defense and that not all persons will deal with threat in the same way. Eriksen (**44**) with a heterogeneous sample of psychiatric patients, found a linear relationship between recognition thresholds for pictures representing aggressive behavior, and ratings of TAT stories for n aggression. Sensitization for the aggressive pictures corresponded with the expression of aggressive content in the stories, and perceptual defense (high thresholds) was coupled with minimal aggressive content, blocking, and incoherent and unelaborated stories when the cards were suggestive of aggressive interpretation. In another study, the same experimenter (**43**) found that disturbance in associating to aggressive, succorant, and homosexual words was positively related to recognition thresholds for scenes depicting people in the act of expressing or gratifying the corresponding needs. Evidence of a similar kind has been presented by Lazarus, Eriksen, and Fonda (**74**) who distinguished patients using "intellectualizing" mechanisms from those using "repressive" mechanisms, and found that the former perceived material significantly more accurately than did the latter. A further experiment by Eriksen (**45**) used groups of *Ss* scoring at the extremes of a measure testing for their recall of completed and incompleted tasks. Perceptual defense was found only in *Ss* who had previously shown an avoid-

ance type of defense in the memory test. This resembles the Postman and Solomon (**103**) finding in which some *Ss* showed relatively high thresholds for words which are associciated with a failure experience, and others showed relatively low thresholds for the same words associated with success experiences. As Eriksen (**46**) points out, an explanation in terms of degrees of familiarity with different words is not appropriate to the data of these two experiments. Furthermore, the response-suppression argument (**126**) is met by them also, since in these experiments the perceptual stimuli were free of social taboos. A recent experiment by Eriksen and Browne (**49**) used groups of *Ss* respectively high and low on the psychasthenia scale of the MMPI. After an experimentally produced failure experience, involving exposure to a list of words, recognition thresholds for the failure-related words and a control series of neutral words were measured. A reduced threshold for the failure words, relative to the neutral words, was found, but there was less reduction for the low psychasthenia group. The significant interaction was regarded as evidence for perceptual defense, which is interpreted in terms of principles derived from punishment and avoidance conditioning.

The finding of McGinnies (**86**) that measurable autonomic reactions to emotionally loaded stimulus material occur at subthreshold exposures was challenged by Aronfreed, Messick, and Diggory (**3**) who found an increase in GSR at the stage of recognition of unpleasant, as contrasted with neutral and pleasant, words. There was no significant difference between thresholds for the latter two classes of words, and hence some evidence is provided for the notion of perceptual defense, but none for perceptual sensitization. In contrast with the findings of some other experimenters (**102, 107**) there were no significant differences between the mean thresholds of "informed" groups ("set" for the type of stimulus to be presented) and "uninformed" groups. Goodstein (**55**) has pointed out that Aronfreed, Messick, and Diggory failed to equate their stimulus words for frequency, and that this variable could have determined the results. Using picture material of aggressive and neutral content, Stein (**119**) studied the responses of a sample of neurotic patients. It was found that these could be classified as either "defenders" or "sensitizers," depending on whether their thresholds for the aggressive material were respectively above or below their thresholds for the neutral material. Subsequent tests established the reliability of the measures and demonstrated the consistency of these patients in adopting one or other of these types of mechanism.

Numerous experiments over the past two years have reported contradictory results and have reflected differing theoretical predilections. De Lucia and Stagner (**40**) report that word-recognition time is clearly affected by two sets of determinants: frequency of usage and emotion-arousing value. It is suggested that future work could usefully aim at relating each of these more effectively to specific personalities. Reece (**111**) obtains results enabling him to conclude that deductions based upon the principles of reward learning theory can effectively predict differences in visual recognition thresholds. Kurland (**67**) using auditory presentation of emotional words, failed to find any difference between the recognition thresholds of obsessive-compulsive

and hysteria patients respectively, a result which would not be predicted from Eriksen's (**45**) hypothesis. Kurland also found that the combined patient groups perceived the emotional words at significantly lower thresholds than did the normal *Ss*, a further discovery inconsistent with much other work.

Another challenging finding is that of Bitterman and Kniffin (**11**) who tested undergraduate women for their recognition of neutral and taboo words, and who also administered to the *Ss* the MMPI and the Taylor Manifest Anxiety Scale. No significant relation between anxiety level and recognition threshold was found. The difference between thresholds for neutral and taboo words was significant, but this difference correlated positively with the *Pd* score of the MMPI, and was unrelated to *K*, *Hy*, *Sc*, or Anxiety. Concluding that the differences in threshold can be better understood in terms of differential readiness to report rather than in terms of perceptual distortion, the authors question an interpretation by McGinnies and Sherman (**88**) of this kind of data. The latter writers assumed that taboo words serve as signals of punishment, and therefore become cues for eliciting anxiety which engenders avoidance reactions. This situation elevates recognition thresholds and may persist long enough to interfere with the perception of subsequent neutral words. Chodorkoff (**36**) has shown that this interpretation is not assailed by the evidence of Bitterman and Kniffin, since the latter writers measured the general level of anxiety of the *Ss*, rather than the crucial variable, i.e., the degree of anxiety which each word is able to elicit. His own work (**35**) supports the view that high- and low-anxiety groups would show differences in recognition thresholds providing that personally relevant stimuli are selected for each *S*. Further work by Chodorkoff (**37**) has led to a revised theoretical position which takes account of new evidence. Word-association time (assumed to measure degree of threat) is unrelated to individual defensiveness as measured by the difference between recognition thresholds for neutral and threatening words, but is related to the absolute value of the perceptual measure; i.e., the degree of deviation of the threshold for the critical words, either above or below that for the neutral words, regardless of sign. Implying that the word-association measure is related to the extent of the perceptual reaction but not to its direction, the data provide further evidence consistent with the view that there are individual differences in the choice of either "vigilant" or "defensive" reactions to threatening stimulation.

Osler and Lewinsohn (**96**), using a carefully controlled stimulus-matching procedure, found that the thresholds for unacceptable words were lower than for acceptable words, a result which runs counter to the majority of previous findings. The data are interpreted as implying that anxiety is associated with greater "vigilance." Neel (**93**) showed tachistoscopically, to various groups of female *Ss*, pictures of persons engaged in various sexual and aggressive activities, and used a multiple-choice situation to obtain responses. Women judged as lacking conflict in the areas of sex and aggression showed "vigilance" to stimuli related to mild sexual behavior, and also revealed "repression" in response to stimuli related to directly sexual situations. There was also in this group "sensitivity" to directly hostile situations

and "avoidance" of stimuli related to mild aggression. The sex-conflict group reacted similarly but less consistently, and the aggression-conflict group tended to "avoid" recognition of all aggressive situations. Kleinman (**66**), measuring changes in auditory perceptual thresholds in cases of psychogenic deafness as compared with cases of organic deafness, also found results consistent with the hypothesis of perceptual defense.

Among the most recent clinically oriented researches which support the notion of a "mechanism" of perceptual defense are the ingenious experiments of Blum. In the first of these (**14**), the Blacky pictures were tachistoscopically presented before and after a situation in which feelings of psychosexual conflict were aroused in the *Ss*. The traumatic picture was selected by the latter as having "stood out the most" significantly more often on the second run than on the first, despite the facts, first, that both series were flashed at subthreshold speeds and, secondly, that there was no conscious recognition of the pictures on either set of trials. This is interpreted as vigilance, at an unconscious level, to cues relevant to the threatening impulses of sex and aggression. With increased exposure times and instructions to locate a particular picture, attempt was made "to bring the ego into play." In this ego-involving situation, significantly fewer correct locations were made of the traumatic than the neutral picture, thus indicating perceptual defense.

Blum's second experiment (**15**) also presented certain Blacky pictures tachistoscopically and was meticulously controlled for the variables of selective verbal report, familiarity, set, and antecedent conditions. Four pictures were simultaneously flashed on a screen at a speed too brief for recognition. Judgments as to which pictures were being shown were nevertheless required. A control condition was supplied by the fact, unknown to the *Ss*, that of the 11 Blacky pictures familiar to them, only 4 were presented. Responses were classified in relation (*a*) to whether the pictures mentioned by *S* were present or absent, and (*b*) to the presence or absence in *S* of conflict plus repression, as independently measured on the Blacky dimensions. To a highly significant degree, *Ss* avoided calling the names of pictures relating to their own conflicts and repressions, but only when these pictures were (subliminally) present. No such avoidance behavior (relative to the pictures which were "neutral" for them) was shown toward the pictures which were not presented. The faults in experimental design which Postman et al. (**107**) showed to be present in early work on perceptual defense are avoided in this study, it is claimed, and furthermore, the results are not easy to interpret within the framework of the Bruner and Postman hypothesis theory of perception.

Blum's work has been extended by Nelson (**94**), with application to the specific personality dynamics of the individual. Finding that the individual perceives in accordance with his areas of high and low conflict and his defense preferences on a variety of psychosexual dimensions, Nelson emphasizes the value of psychoanalytic theory as a basis for research upon perceptual vigilance and defense.

Discussion

A proportion of the literature reviewed above seeks to reduce the phenomenon of defense to simpler and more familiar principles, to minimize its significance as a special

problem and to challenge its interpretation in terms foreign to the existing body of general psychological theory. Such attempts are plausible with respect to some specific results, but are probably inadequate as a basis for explaining all available data. As pointed out by Postman, Bruner, and McGinnies (**101**), the same problem exists in the phenomena of hysterical and hypnotically induced blindness. Data of this kind cannot be readily explained by the kind of arguments directed at some laboratory studies of perceptual defense. Conservatively biased criticism has indeed served a useful purpose, as indicated above, but it must now be admitted that Eriksen and Browne (**49**) have correctly summed up the position in stating that "a firm body of experimental support for such a phenomenon has remained untouched by criticism."

Among the experimentally oriented clinicians who have increasingly entered this field, the opinion has grown that perceptual defense is not a phenomenon pointing to some underlying general law. Some general laws (e.g., sensitivity to frequently appearing words) may influence its manifestation under some conditions, but for a clear demonstration of its appearance with the factors of set, selective report, and frequency controlled, it is necessary to design the experiment to take account of certain critical individual differences.

On the other hand, perceptual theory has the task of accounting for perceptual phenomena, regardless of whether individual differences are systematically included or minimized in the experimental design. The search for a general explanation appeared at a relatively early stage in the controversy about "perceptual defense." Bruner and Postman (**28**) were responsible for introducing the notion that recognition need not necessarily be defined as coincident with veridical report and that there may be a hierarchy of response thresholds to a given stimulus. Some of these, such as an affective-avoidance reaction, may well be tripped off prior to the threshold for veridical report. Testing an explicit hypothesis along these lines, McGinnies (**86**) gave some experimental support to this kind of theorizing. Subsequently, McCleary and Lazarus (**73, 82**) reported evidence for a phenomenon termed "subception" or discriminative autonomic response to subliminal stimulation. Support for this finding has come from Taylor (**120**) and from Rubenfeld, Lowenfeld, and Guthrie (**78, 113**). A challenge to the interpretation of such data as "discrimination without awareness" has been presented by Bricker and Chapanis (**16**), Howes (**60**), Murdock (**91**), Lysak (**80**), Eriksen (**47**), and Voor (**124**). Such an interpretation, however, is not crucial to the hypothesis of Bruner and Postman. The affective-avoidance reaction could be mediated by a conscious, though partial, recognition, equally as well as by an "unconscious awareness." In either case, the work on so-called "subception" is consistent with their view which still stands as a plausible explanation of much work on perceptual defense.

An alternative view, supported by experimental evidence, is offered by Hochberg, Haber, and Ryan (**59**). It is supposed that during the interval between stimulation and report, a rapid sequence of events may occur, involving, for example, an autonomic response. This may be so strong, relative to the tender, newborn memory trace, as to disrupt the latter and thus prevent recognition and recall of the briefly presented material.

Explanations such as the two cited above have the appeal of elegant simplicity. Both are susceptible of application to data concerning differences in the type of material for which defense occurs and in the type of *S* who shows this behavior. These interpretations relate strictly to perceptual defense, however, and make no attempt to subsume either the "vigilance" or the "sensitization" effects. It would seem uneconomical to have two or three different principles to explain classes of perceptual phenomena which seem closely related. Further research aimed at clarifying the interrelationships of the three types of phenomena, and at exploring the possibility of a unitary underlying principle would seem to be required.

The Definition of Perception

Review of the present field would be incomplete without some reference to a major issue; that is, the definition of *perception* and the specification of the level at which need states, personality factors, past experiences, and present expectations operate upon the experience reported by *S*. The great majority of experiments reviewed above have introduced a measure of ambiguity into the stimulus situation. By brief or unclear exposure of the material, or by a delay between stimulus and response, maximum play has been given to subjective factors. Can the operation of these be described in terms of thinking, imagining, or problem solving rather than as part of a broadly defined perceptual process?

Where *S* is provided with fuller information, as in "conventional" psychophysics (so styled to distinguish it from some studies reviewed above), his behavior tends to support the view that the percept is stimulus-bound (**108**). The social or clinical psychologist, on the other hand, finds it useful to broaden his definition. Bruner (**21**), for example, justifies the experimenter's use of ambiguity by noting that: " . . . most complex perception, particularly in our social lives, is dependent upon the integration of information of a far less reliable kind than we normally provide in a tachistoscope at rapid exposure."

The question involves more than semantic convenience. To some, it is a central theoretical issue. An instance is provided by Wallach's (**125**) distinction between a sensorily determined perceptual experience and a recalled trace complex which gives identity and "total meaning" to the experience. Acceptance of this premise could lead to the argument that need operates upon the recalled trace complex rather than upon the perceptual experience. But the validity of the premise is open to question, as shown, for example, by the experiment of Hastorf (**57**) who found that the apparent distance of an object depended on the degree of assumed size attributed to it. That is to say, identification of the object was a necessary condition for the "primary" perceptual experience of distance. Pratt (**108**) argues that Hastorf's results are explained by a shift in the judgmental frame of reference, and Prentice (**109**) questions whether the location of the object really *looked* different under the different experimental conditions. Such skepticism appears rather forced when it is noted that Hastorf dealt carefully with this point in his report (**57**, pp. 208–209 and pp. 212–213) and concluded on the basis of both quantitative and qualitative evidence that the judgments of the subjects "had very definite perceptual aspects and were not purely intellectual in nature."

More recently, Bruner and Minturn (31) have pointed out that the act of identification can modify the primitive perceptual organization of the field. Their experiment showed unequivocally that, for their *Ss*, closure was not a self-determining and "pure" perceptual or stimulus process, but operated differentially with respect to the identification given the object. A broken letter "B" was recognized as a B when the subjects expected to see a letter, and was recognized as the figure 13 when they were expecting to see a number. Work of this kind gives us good reason to believe that "perception" on the one hand, and "identification," or "recognition" on the other hand, while analytically separable under some conditions, cannot be distinguished under others, and hence the distinction is theoretically untenable. If this be accepted, then certainly it cannot be agreed that need and perception relationships are "so hard to demonstrate" (109), and it should probably be disputed also that such relationships lack importance and generality (123). On the other hand, it must be recognized that the controversy is far from its final resolution. While further discussion would take us beyond the limits set for the present review, it might at least be concluded that progress is needed not only in need and perception research but also in memory and concept formation before adequate perspective can be reached (cf. 22, 32).

Summary and Conclusions

The present review was prompted by the divergent evaluations of several commentators upon research in the field of affective processes in perception. Four areas were selected in which activity has been strong for several years past. Studies of size judgment provide a category defined by the general nature of the dependent variable. Design and technique in such experiments has progressively improved. Several findings indicate that certain motivational states are determinants of size judgment, and this is true of some recent work as well as of the earlier studies. A satisfactory theory to account for the correlation is lacking. Physiological need in relation to perception is the second area considered. From the several studies reviewed it is concluded that the weight of evidence favors the hypothesis that need is a determinant of perception, but that only a beginning has been made in the search for reasons to explain this relationship. A third area comprises a group of studies on "selective sensitization" for stimuli representing positive values. Such a phenomenon appears to be well established, and not wholly due to artifacts of experimental procedure. Adequate specification of the conditions under which sensitization occurs waits upon future activity. The final, and most prolific, area dealt with concerns reactions to stimuli presumed to be noxious or threatening to *Ss*. Here are reviewed a considerable number of studies attempting to demonstrate or deny the phenomenon known as "perceptual defense." Challenging findings are presented by clinically oriented studies involving individual differences and also by experiments related to theory of a more generalized type. The problem of discovering the mediating mechanisms responsible for the perceptual-motivational correlation has stimulated some useful research, as witnessed, for example, by the work on "subception," and has also provoked some thinking which will direct future investigations. A final section is included which deals

with some contrasting ways of defining the term *perception*.

It can be concluded that studies on perception in relation to various affective processes have been amply successful in the raising of important problems and the setting of useful directions for future work. Integration of the complex and oft-seeming contradictory body of data is much needed, but some progress is being made in this direction. Future work requires renewed sifting of the evidence, replication of studies, particularly those with conflicting results, continued improvement in methodology, a sense of direction towards theoretical objectives, recognition of work in related areas, and at least some coordination of effort between the widely ramifying branches of the field.

REFERENCES

1. Adams, J., & Brown, D. R. Values, word frequencies and perception. *Psychol. Rev.*, 1953, **60**, 50–54.
2. Allport, F. H. *Theories of perception and the concept of structure.* New York: Wiley, 1955.
3. Aronfreed, J. M., Messick, S. A., & Diggory, J. C. Re-examining emotionality and perceptual defense. *J. Pers.*, 1953, **21**, 517–528.
4. Ashley, W. R., Harper, R. S., & Runyon, D. L. The perceived size of coins in normal and hypnotically induced economic states. *Amer. J. Psychol.*, 1951, **64**, 564–572.
5. Atkinson, J. W., & McClelland, D. C. The projective expression of needs: II. The effect of different intensities of the hunger drive on thematic apperception. *J. exp. Psychol.*, 1948, **38**, 643–658.
6. Ayllon, T., & Sommer, R. Autism, emphasis and figure-ground perception. *J. Psychol.*, 1956, **41**, 163–176.
7. Beach, F. A. Body chemistry and perception. In R. R. Blake & G. V. Ramsey (Eds.), *Perception: an approach to personality.* New York: Ronald, 1951. Pp. 56–94.
8. Beams, H. L. Affectivity as a factor in the apparent size of pictured food objects. *J. exp. Psychol.*, 1954, **47**, 197–200.
9. Beier, E. G., & Cowen, E. L. A further investigation of the influence of "threat-expectancy" on perception. *J. Pers.*, 1953, **22**, 254–257.
10. Bevan, W., Jr., & Dukes, W. F. Value and the Weber Constant in the perception of distance. *Amer. J. Psychol.*, 1951, **64**, 580–584.
11. Bitterman, M. E., & Kniffin, C. W. Manifest anxiety and "perceptual defense." *J. abnorm. soc. Psychol.*, 1953, **48**, 248–253.
12. Blake, R. R., & Vanderplas, J. M. The effect of prerecognition hypotheses on veridical recognition thresholds in auditory perception. *J. Pers.*, 1950, **19**, 95–115.
13. Blake, R. R., & Ramsey, G. V. *Perception: an approach to personality.* New York: Ronald, 1951.
14. Blum, G. S. An experimental reunion of psychoanalytic theory with perceptual vigilance and defense. *J. abnorm. soc. Psychol.*, 1954, **49**, 94–98.
15. Blum, G. S. Perceptual defense revisited. *J. abnorm. soc. Psychol.*, 1955, **51**, 24–29.
16. Bricker, P. D., & Chapanis, A. Do incorrectly perceived tachistoscopic stimuli convey some information? *Psychol. Rev.*, 1953, **60**, 181–188.
17. Brown, D. R., & Adams, J. Word frequency and the measurement of value areas. *J. abnorm. soc. Psychol.*, 1954, **49**, 427–430.
18. Brozek, J., Guetzkow, H., & Baldwin, M. G. A quantitative study of perception and association in experimental semistarvation. *J. Pers.*, 1951, **19**, 245–264.
19. Bruner, J. S. Perceptual theory and the Rorschach Test. *J. Pers.*, 1948, **17**, 157–168.
20. Bruner, J. S. One kind of perception: A reply to Professor Luchins. *Psychol. Rev.*, 1951, **58**, 306–312.
21. Bruner, J. S. Personality dynamics and the process of perceiving. In R. R. Blake & G. V. Ramsey (Eds.), *Perception: an approach to personality.* New York: Ronald, 1951. Pp. 121–147.

22. BRUNER, J. S. On perceptual readiness. *Psychol. Rev.*, in press.
23. BRUNER, J. S., & POSTMAN, L. Emotional selectivity in perception and reaction. *J. Pers.*, 1947, **16**, 69–77.
24. BRUNER, J. S., & POSTMAN, L. Tension and tension release as organizing factors in perception. *J. Pers.*, 1947, **15**, 300–308.
25. BRUNER, J. S., & GOODMAN, CECILE C. Value and need as organizing factors in perception. *J. abnorm. soc. Psychol.*, 1947, **42**, 33–44.
26. BRUNER, J. S., & POSTMAN, L. An approach to social perception. In W. Dennis (Ed.), *Current trends in social psychology*. Pittsburgh: Univer. of Pittsburgh Press, 1948. Pp. 71–118.
27. BRUNER, J. S., & POSTMAN, L. Symbolic value as an organizing factor in perception. *J. soc. Psychol.*, 1948, **27**, 203–208.
28. BRUNER, J. S., & POSTMAN, L. Perception, cognition and behavior. *J. Pers.*, 1949, **18**, 14–31.
29. BRUNER, J. S., & KRECH, D. (Eds.) *Perception and personality: A symposium*. Durham: Duke Univer. Press, 1950.
30. BRUNER, J. S., & RODRIGUES, J. S. Some determinants of apparent size. *J. abnorm. soc. Psychol.*, 1953, **48**, 17–24.
31. BRUNER, J. S., & MINTURN, A. L. Perceptual identification and perceptual organization. *J. gen. Psychol.*, 1955, **53**, 21–28.
32. BRUNER, J. S., GOODNOW, JACQUELINE J., & AUSTIN, G. A. *A study of thinking*. New York: Wiley, 1956.
33. CARPENTER, B., WIENER, M., & CARPENTER, JANETH T. Predictability of perceptual defense behavior. *J. abnorm. soc. Psychol.*, 1956, **52**, 380–383.
34. CARTER, L. F., & SCHOOLER, K. Value need and other factors in perception. *Psychol. Rev.*, 1949, **56**, 200–207.
35. CHODORKOFF, B. Self perception, perceptual defense and adjustment. *J. abnorm. soc. Psychol.*, 1954, **49**, 508–512.
36. CHODORKOFF, B. A note on Bitterman & Kniffin's "Manifest anxiety and perceptual defense." *J. abnorm. soc. Psychol.*, 1955, **50**, 144.
37. CHODORKOFF, B. Anxiety, threat and defensive reactions. *J. gen. Psychol.*, 1956, **54**, 191–196.
38. COWEN, E. L., & BEIER, E. G. The influence of threat-expectancy on perception. *J. Pers.*, 1950, **19**, 85–94.
39. COWEN, E. L., & BEIER, E. G. Threat-expectancy, word frequencies and perceptual prerecognition hypotheses. *J. abnorm. soc. Psychol.*, 1954, **49**, 178–182.
40. DE LUCIA, J. L., & STAGNER, R. Emotional vs. frequency factors in word recognition time and association time. *J. Pers.*, 1954, **22**, 299–309.
41. DUKES, W. F., & BEVAN, W., JR. Accentuation and response variability in the perception of personally relevant objects. *J. Pers.*, 1952, **20**, 457–465.
42. DUKES, W. F., & BEVAN, W., JR. Size estimation and monetary value: a correlation. *J. Psychol.*, 1952, **34**, 43–53.
43. ERIKSEN, C. W. Perceptual defense as a function of unacceptable needs. *J. abnorm. soc. Psychol.*, 1951, **46**, 557–564.
44. ERIKSEN, C. W. Some implications for TAT interpretation arising from need and perception experiments. *J. Pers.*, 1951, **19**, 282–288.
45. ERIKSEN, C. W. Defense against ego threat in memory and perception. *J. abnorm. soc. Psychol.*, 1952, **47**, 230–235.
46. ERIKSEN, C. W. The case for perceptual defense. *Psychol. Rev.*, 1954, **61**, 175–182.
47. ERIKSEN, C. W. Subception: fact or artifact? *Psychol. Rev.*, 1956, **63**, 74–80.
48. ERIKSEN, C. W., & LAZARUS, R. S. Perceptual defense and projective tests. *J. abnorm. soc. Psychol.*, 1952, **47**, 302–308.
49. ERIKSEN, C. W., & BROWNE, C. T. An experimental and theoretical analysis of perceptual defense. *J. abnorm. soc. Psychol.*, 1956, **52**, 224–230.
50. FREEMAN, J. T. Set or perceptual defense. *J. exp. Psychol.*, 1954, **48**, 283–288.
51. FREEMAN, J. T. Set versus perceptual defense: a confirmation. *J. abnorm. soc. Psychol.*, 1955, **51**, 710–712.
52. GILCHRIST, J. C., & NESBERG, L. S. Need and perceptual change in need-related objects. *J. exp. Psychol.*, 1952, **44**, 369–377.
53. GILCHRIST, J. C., LUDEMAN, J. F., & LYSAK, W. Values as determinants of word recognition thresholds. *J. abnorm. soc. Psychol.*, 1954, **49**, 423–426.
54. GOLLIN, E. S., & BARON, A. Response consistency in perception and retention. *J. exp. Psychol.*, 1954, **47**, 259–262.

55. GOODSTEIN, L. D. Affective tone and visual recognition thresholds. *J. abnorm. soc. Psychol.*, 1954, **49**, 443–444.
56. HAIGH, G. V., & FISKE, D. W. Corroboration of personal values as selective factors in perception. *J. abnorm. soc. Psychol.*, 1952, **47**, 394–398.
57. HASTORF, A. H. The influence of suggestion on the relation between stimulus size and perceived distance. *J. Psychol.*, 1950, **29**, 195–217.
58. HENLE, MARY. Some effects of motivational processes on cognition. *Psychol. Rev.*, 1955, **62**, 423–432.
59. HOCHBERG, J. E., HABER, S. L., & RYAN, T. A. "Perceptual defense" as an interference phenomenon. *Percept. Mot. Skills*, 1955, **5**, 15–17.
60. HOWES, D. A statistical theory of the phenomenon of subception. *Psychol. Rev.*, 1954, **61**, 98–110.
61. HOWES, D. H., & SOLOMON, R. L. Visual duration threshold as a function of word-probability. *J. exp. Psychol.*, 1951, **41**, 401–410.
62. HOWIE, D. Perceptual defense. *Psychol. Rev.*, 1952, **59**, 308–315.
63. KLEIN, G. S. Need and regulation. In M. R. Jones (Ed.), *Nebraska symposium on motivation*. Lincoln: Univer. of Nebraska Press, 1954. Pp. 224–274.
64. KLEIN, G. S., & SCHLESINGER, H. Where is the perceiver in perceptual theory? *J. Pers.*, 1949, **18**, 32–47.
65. KLEIN, G. S., SCHLESINGER, H. J., & MEISTER, D. E. The effect of personal values on perception: an experimental critique. *Psychol. Rev.*, 1951, **58**, 96–112.
66. KLEINMAN, M. L. Psychogenic deafness and perceptual defense. *Amer. Psychologist*, 1954, **9**, 406. (Abstract)
67. KURLAND, S. H. The lack of generality in defense mechanisms as indicated in auditory perception. *J. abnorm. soc. Psychol.*, 1954, **49**, 173–177.
68. LACY, O. W., LEWINGER, N., & ADAMSON, J. F. Foreknowledge as a factor affecting perceptual defense and alertness. *J. exp. Psychol.*, 1953, **45**, 169–174.
69. LAMBERT, W. W., SOLOMON, R. L., & WATSON, P. D. Reinforcement and extinction as factors in size estimation. *J. exp. Psychol.*, 1949, **39**, 637–641.
70. LAMBERT, W. W., & LAMBERT, ELIZABETH C. Some indirect effects of reward on children's size estimations. *J. abnorm. soc. Psychol.*, 1953, **48**, 507–510.
71. LAZARUS, R. S. Ambiguity and nonambiguity in projective testing. *J. abnorm. soc. Psychol.*, 1953, **48**, 443–445.
72. LAZARUS, R. S. Is there a mechanism of perceptual defense? A reply to Postman, Bronson and Gropper. *J. abnorm. soc. Psychol.*, 1954, **49**, 396–398.
73. LAZARUS, R. S., & MCCLEARY, R. A. Autonomic discrimination without awareness: a study of subception. *Psychol. Rev.*, 1951, **58**, 113–122.
74. LAZARUS, R. S., ERIKSEN, C. W., & FONDA, C. P. Personality dynamics and auditory perceptual recognition. *J. Pers.*, 1951, **19**, 471–482.
75. LAZARUS, R. S., YOUSEM, H., & ARENBERG, D. Hunger and perception. *J. Pers.*, 1953, **21**, 312–328.
76. LEVINE, R., CHEIN, I., & MURPHY, G. The relation of the intensity of a need to the amount of perceptual distortion: a preliminary report. *J. Psychol.*, 1942, **13**, 283–293.
77. LINDNER, H. Sexual responsiveness to perceptual tests in a group of sexual offenders. *J. Pers.*, 1953, **21**, 364–375.
78. LOWENFELD, J., RUBENFELD, S., & GUTHRIE, G. M. Verbal inhibition in subception. *J. gen. Psychol.*, 1956, **54**, 171–176.
79. LUCHINS, A. S. An evaluation of some current criticisms of Gestalt work on perception. *Psychol. Rev.*, 1951, **58**, 69–95.
80. LYSAK, W. The effects of punishment upon syllable recognition thresholds. *J. exp. Psychol.*, 1954, **47**, 343–350.
81. LYSAK, W., & GILCHRIST, J. C. Value, equivocality, and goal availability as determinants of size judgments. *J. Pers.*, 1955, **23**, 500–501. (Abstract)
82. MCCLEARY, R. A., & LAZARUS, R. S. Autonomic discrimination without awareness: an interim report. *J. Pers.*, 1949, **18**, 171–179.
83. MCCLELLAND, D. C. *Personality*. New York: William Sloane Associates, 1951.
84. MCCLELLAND, D. C., & ATKINSON, J. W. The projective expression of needs: I. The effects of different intensities of the hunger drive on perception. *J. Psychol.*, 1948, **25**, 205–222.
85. MCCLELLAND, D. C., & LIBERMAN, A. M. The effect of need for achievement on recognition of need-related words. *J. Pers.*, 1949, **18**, 236–251.
86. MCGINNIES, E. Emotionality and per-

ceptual defense. *Psychol. Rev.*, 1949, **56**, 244–251.

87. McGINNIES, E., & BOWLES, W. Personal values as determinants of perceptual fixation. *J. Pers.*, 1949, **18**, 224–235.
88. McGINNIES, E., & SHERMAN, H. Generalization of perceptual defense. *J. abnorm. soc. Psychol.*, 1952, **47**, 81–85.
89. MAUSNER, B., & SIEGEL, A. The effect of variation in "value" on perceptual thresholds. *J. abnorm. soc. Psychol.*, 1950, **45**, 760–763.
90. MICHAUX, W. Schizophrenic apperception as a function of hunger. *J. abnorm. soc. Psychol.*, 1955, **50**, 53–58.
91. MURDOCK, B. B., JR. Perceptual defense and threshold measurements. *J. Pers.*, 1954, **22**, 565–571.
92. MURPHY, G. Affect and perceptual learning. *Psychol. Rev.*, 1956, **63**, 1–15.
93. NEEL, ANN F. Conflict, recognition time and defensive behavior. *Amer. Psychologist*, 1954, **9**, 437. (Abstract)
94. NELSON, S. E. Psychosexual conflicts and defenses in visual perception. *J. abnorm. soc. Psychol.*, 1955, **51**, 427–433.
95. NEWTON, K. R. A note on visual recognition thresholds. *J. abnorm. soc. Psychol.*, 1955, **51**, 709–710.
96. OSLER, SONIA F., & LEWINSOHN, P. M. The relation between manifest anxiety and perceptual defense. *Amer. Psychologist*, 1954, **9**, 446. (Abstract)
97. PASTORE, N. Need as a determinant of perception. *J. Psychol.*, 1949, **28**, 457–475.
98. POSTMAN, L. Toward a general theory of cognition. In J. H. Rohrer & M. Sherif (Eds.), *Social psychology at the crossroads*. New York: Harper, 1951. Pp. 242–272.
99. POSTMAN, L. The experimental analysis of motivational factors in perception. In J. S. Brown, H. F. Harlow, L. Postman, et al., *Current theory and research in motivation: a symposium*. Lincoln: Univer. of Nebraska Press, 1953. Pp. 59–108.
100. POSTMAN, L. On the problem of perceptual defense. *Psychol. Rev.*, 1953, **60**, 298–306.
101. POSTMAN, L., BRUNER, J. S., & McGINNIES, E. Personal values as selective factors in perception. *J. abnorm. soc. Psychol.*, 1948, **43**, 142–154.
102. POSTMAN, L., & BRUNER, J. S. Multiplicity of set as a determinant of perceptual organization. *J. exp. Psychol.*, 1949, **39**, 369–377.
103. POSTMAN, L., & SOLOMON, R. L. Perceptual sensitivity to completed and incompleted tasks. *J. Pers.*, 1950, **18**, 347–357.
104. POSTMAN, L., & SCHNEIDER, B. Personal values, visual recognition and recall. *Psychol. Rev.*, 1951, **58**, 271–284.
105. POSTMAN, L., & CRUTCHFIELD, R. S. The interaction of need, set and stimulus structure in a cognitive task. *Amer. J. Psychol.*, 1952, **65**, 196–217.
106. POSTMAN, L., & BROWN, D. R. The perceptual consequences of success and failure. *J. abnorm. soc. Psychol.*, 1952, **47**, 213–221.
107. POSTMAN, L., BRONSON, W. C., & GROPPER, G. L. Is there a mechanism of perceptual defense? *J. abnorm. soc. Psychol.*, 1953, **48**, 215–225.
108. PRATT, C. C. The role of past experience in visual perception. *J. Psychol.*, 1950, **30**, 85–107.
109. PRENTICE, W. C. H. "Functionalism" in perception. *Psychol. Rev.*, 1956, **63**, 29–38.
110. PROSHANSKY, H., & MURPHY, G. The effects of reward and punishment on perception. *J. Psychol.*, 1942, **13**, 295–305.
111. REECE, M. M. The effect of shock on recognition thresholds. *J. abnorm. soc. Psychol.*, 1954, **49**, 165–172.
112. ROSEN, A. C. Change in perceptual threshold as a protective function of the organism. *J. Pers.*, 1954, **23**, 182–194.
113. RUBENFELD, S., LOWENFELD, J., & GUTHRIE, G. M. Stimulus generalization in subception. *J. gen. Psychol.*, 1956, **54**, 177–182.
114. SANFORD, R. N. The effect of abstinence from food upon imaginal processes. *J. Psychol.*, 1936, **2**, 129–136.
115. SCHAFER, E., & MURPHY, G. The role of autism in a visual figure-ground relationship. *J. exp. Psychol.*, 1943, **32**, 335–343.
116. SOLLEY, C. M., & LEE, R. Perceived size: closure versus symbolic value. *Amer. J. Psychol.*, 1955, **68**, 142–144.
117. SOLOMON, R. L., & HOWES, D. W. Word frequency, personal values and visual deviation thresholds. *Psychol. Rev.*, 1951, **58**, 256–270.
118. SOLOMON, R. L., & POSTMAN, L. Frequency of usage as a determinant of recognition threshold for words. *J. exp. Psychol.*, 1952, **43**, 195–202.

119. STEIN, K. B. Perceptual defense and perceptual sensitization under neutral and involved conditions. *J. Pers.*, 1953, **21**, 467–478.
120. TAYLOR, F. W. R. The discrimination of subliminal visual stimuli. *Canad. J. Psychol.*, 1953, **7**, 12–20.
121. TAYLOR, JANET A. Physiological need, set, and visual duration threshold. *J. abnorm. soc. Psychol.*, 1956, **52**, 96–99.
122. VANDERPLAS, J. M., & BLAKE, R. R. Selective sensitization in auditory perception. *J. Pers.*, 1949, **18**, 252–266.
123. VERNON, MAGDALEN D. The functions of schemata in perceiving. *Psychol. Rev.*, 1955, **62**, 180–192.
124. VOOR, J. H. Subliminal perception and subception. *J. Psychol.*, 1956, **41**, 437–458.
125. WALLACH, H. Some considerations concerning the relation between perception and cognition. *J. Pers.*, 1949, **18**, 6–13.
126. WHITTAKER, EDNA M., GILCHRIST, J. C., & FISCHER, JEAN W. Perceptual defense or response suppression? *J. abnorm. soc. Psychol.*, 1952, **47**, 615–623.
127. WIENER, M. Word frequency or motivation in perceptual defense. *J. abnorm. soc. Psychol.*, 1955, **51**, 214–218.
128. WISPÉ, L. G. Physiological need, verbal frequency and word association. *J. abnorm. soc. Psychol.*, 1954, **49**, 229–234.
129. WISPÉ, L. G., & DRAMBAREAN, N. C. Physiological need, word frequency and visual duration thresholds. *J. exp. Psychol.*, 1953, **46**, 25–31.

Received August 2, 1956.

PSYCHOLOGICAL BULLETIN
Vol. 54, No. 2, 1957

ADDITIONAL "POST-MORTEM" TESTS OF EXPERIMENTAL COMPARISONS

JULIAN C. STANLEY
University of Wisconsin[1]

McHugh and Ellis (3) furnish an illustration of Scheffé's (5) method for judging all contrasts in the analysis of variance. It seems desirable to emphasize here that Scheffé's procedure, being quite general, is more conservative than necessary for most comparisons of interest to psychologists—that is, it then results in too few rejections of the null hypothesis and too wide confidence intervals around the obtained difference between means. Frequently, alternative tests suggested by Tukey[2] and Dunnett (1) will be more powerful.

Scheffé offers a method for testing among means all possible contrasts that were not incorporated explicitly into the experimental design, provided only that the *sum* of the weights assigned to the various means is zero[3] (for planned contrasts, use Student's t). Suppose there are 4 means, as in the McHugh-Ellis analysis. We may arbitrarily take the sum of 10 times the first mean, 3 times the second mean, and 6 times the third, so long as we subtract from this sum $(10+3+6)=19$ times the fourth mean. The variance of this difference among these independent means will be 10^2 times the variance of the first mean plus 3^2 times the variance of the second plus 6^2 times the variance of the third plus 19^2 times the variance of the fourth. Since the random-sampling variance of a mean is σ^2/n and we use the mean square for error, s^2, as an estimate of the population variance for each group, the variance of the weighted composite above will be $s^2(100/n_1+9/n_2+36/n_3+361/n_4)$. In the McHugh-Ellis article, $s^2=31.5$ and each n_i is 12, so the variance of the composite is

$$\frac{31.5(100+9+36+361)}{12}=1328.25.$$

The square root of this, 36.45, is to be compared in some manner with the net difference among the weighted means, $10(105.61)+3(112.27)+6(103.93)-19(114.05)=-150.46$. The ratio of 150.46 to 36.45 is 4.13, which we might naively (and incorrectly) compare with a t with 32 df at, say, the $\alpha=.01$ level of significance for a two-tailed test, which is 2.74. The appropriate comparison is with

$$\sqrt{(k-1)F_{(1-\alpha,\ k-1,\ df_2)}} = \sqrt{(4-1)_{.99}F_{(4-1,\ 32)}} = \sqrt{3(4.46)}=3.66,$$

$k=4$ being the total number of means that might be compared. Since 4.13 exceeds 3.66, the difference is significant beyond the .01 level. The *in*correct .99 confidence interval is $-150.46 \pm 2.74(36.45) = -250.33$ to -50.59,

[1] Postdoctoral fellow in statistics, University of Chicago, 1955–56. Thanks are due James C. Reed for helpful suggestions.

[2] McHugh and Ellis mention Tukey's test in a footnote but do not illustrate it, nor do they emphasize the overgenerality of Scheffé's procedure for most psychological experiments. In a long, mimeographed, undated manuscript entitled "The problem of multiple comparisons" Tukey gives the rationale for his method, which Scheffé (5) discusses.

[3] Even when the over-all F is not significant, one may ascertain precise confidence limits by the Scheffé, Tukey, or Dunnett procedures.

while the correct interval is much wider: $-150.46 \pm 3.66(36.45) = -283.87$ to -17.05. In this experiment one may, knowing 3.66, set up a confidence interval for *any* comparison,

$$\sum_{i=1}^{k} c_i \bar{X}_i,$$

the c_is being fixed weights ($-\infty < c_i < \infty$) such that

$$\sum_{i=1}^{k} c_i = 0.$$

Often, perhaps usually, we make a posteriori contrasts of just two means, $\bar{X}_i$ and $\bar{X}_j$, weighted equally; let $c_i = 1$, $c_j = -1$, and all other cs $= 0$. For this particular type of comparison, and if $n_i = n_j$, Tukey's method is preferable to Scheffé's. As an illustration, consider .99 confidence intervals for $\bar{X}_2 - \bar{X}_1 = 112.27 - 105.61 = 6.66$ from the McHugh-Ellis data. The simple t-test limits are too narrow:

$$6.66 \pm 2.74\sqrt{31.5}\sqrt{1/12+1/12} = 6.66 \pm 6.27.$$

Scheffé's are too broad:

$$6.66 \pm 3.66\sqrt{31.5}\sqrt{1/12+1/12} = 6.66 \pm 8.38.$$

Tukey's are intermediate between the above two:

$$6.66 \pm q\sqrt{31.5}\sqrt{1/12} = 6.66 \pm 7.74,$$

where $q = 4.775$ is the upper 1% point of Studentized range (**4**, p. 177) for a sample of 4 means with 32 *df* for s^2.

Thus when interested only in contrasts of the form $(\bar{X}_i - \bar{X}_j)$, use Tukey's method if the n_is are equal. For differing n_is and/or more complicated contrasts, use Scheffé's procedure, which is completely general within the restriction that the sum of the weights applied to the k means must be zero. All three methods (simple t test, Tukey's, and Scheffé's) give identical results when $k = 2$.

Where a "standard" treatment has been incorporated into the experimental design, as for example when one control group and three different experimental groups are used, we can narrow the confidence interval even more by employing Dunnett's (**1**) tables. Suppose that in the above example Method 2 was the predesignated standard with which each of the three other methods was to be compared. The appropriate .99 confidence interval for the contrast with Method 1 would be $6.66 \pm 3.15\sqrt{31.5}\sqrt{2/12} = 6.66 \pm 7.21$, where 3.15 is the figure obtained from Dunnett's Table 2*b* (**1**, p. 1120) for 3 "treatment" means and 32 *df*.[4] This interval will typically be narrower than limits secured via Tukey's procedure, but wider than those from the simple t test, unless $k = 2$, when all are identical.[5]

The Tukey and Dunnett methods both require that $n_1 = n_2 = \cdots = n_k$. Otherwise, probability values for the confidence intervals are approximate.

Concluding Remarks

The techniques for obtaining confidence intervals described above ap-

[4] He also furnishes tables for one-tailed comparisons, to be used when an explicit hypothesis specifying the direction of deviation of the treatment mean from the standard was stated *before* measures were secured.

[5] By D. B. Duncan's multiple range test (Multiple range and multiple F tests, *Biometrics*, 1955, **11**, 1–42), four of the six possible simple comparisons among the four means are significant at the .01 level. M_4 (114.05) does not differ significantly from M_2 (112.27), nor does M_1 (105.61) from M_3 (103.93).

ply especially to Model I (fixed-effect) designs where levels of the factor tested are unordered. For the fixed effects in "mixed" models, see Scheffé (**6**, p. 33; **7**, pp. 261–263). Kurtz (**2**) outlines a procedure to use when error variances are unequal.

REFERENCES

1. DUNNETT, C. W. A multiple comparison procedure for comparing several treatments with a control. *J. Amer. statist. Ass.*, 1955, **50**, 1096–1121.
2. KURTZ, T. E. An extension of a method of marking multiple comparisons (preliminary report). *Ann. math. Statist.*, 1956, **27**, 547. (Abstract)
3. MCHUGH, R. B., & ELLIS, D. S. The "post-mortem" testing of experimental comparisons. *Psychol. Bull.*, 1955, **52**, 425–428.
4. PEARSON, E. S., & HARTLEY, H. O. (Eds.) *Biometrika tables for statisticians*. Vol. I. Cambridge: Cambridge Univer. Press 1954.
5. SCHEFFÉ, H. A method for judging all contrasts in the analysis of variance. *Biometrika*, 1953, **40**, 87–104.
6. SCHEFFÉ, H. A "mixed model" for the analysis of variance. *Ann. math. Statist.*, 1956, **27**, 23–36.
7. SCHEFFÉ, H. Alternative models for the analysis of variance. *Ann. math. Statist.*, 1956, **27**, 251–271.

Received March 9, 1956.

PSYCHOLOGICAL BULLETIN
Vol. 54, No. 2, 1957

WITHIN-GROUPS CORRELATIONS AND THEIR CORRECTION FOR ATTENUATION

JOHN R. HILLS
Educational Testing Service

Situations frequently occur in which an investigator obtains the correlations between the same two variables in several groups of subjects and wants to obtain some over-all estimate of the degree of correlation between the two variables. For example, one might find the correlations between scores on a measure of Mechanical Information (Me) and scores on a measure of Mathematics Background (MB) for students in several trade and technical schools and want to know a general value expressing the degree of relationship between these two variables.

With respect to the variables being correlated various situations might exist. Several examples are shown in Fig. 1. A set of data such as that in Fig. 1A might occur if students in some schools were extensively trained in mathematics but had little shop experience, while students in other schools received the opposite pattern of training. Note that within the schools (A, B, C, D) the correlations are all positive and about equal in magnitude, but, due to the negative correlation between means, a coefficient obtained by plotting the data from all schools on a single scatter diagram would be negative (the dotted ellipse). Figure 1B represents a situation with a perfect positive correlation between means. The coefficient of correlation obtained in any school, such as G, might be corrected for restriction in range, in this case, to give the value which would be obtained if all data were plotted on a single scatter diagram. In Fig. 1C the correlation between means is about equal to the correlation within the groups, and the dotted ellipse has about the same shape as the solid-line ellipses.

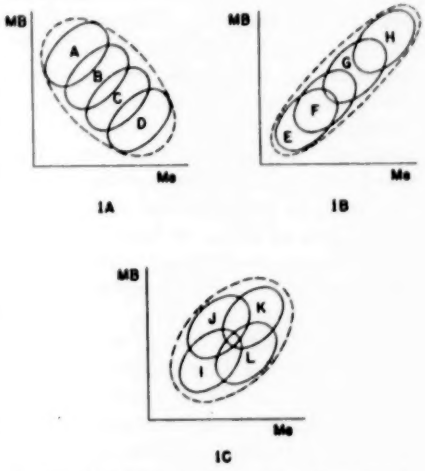

FIG. 1. ILLUSTRATION OF THE INFLUENCE OF CORRELATIONS BETWEEN SUBGROUP MEANS ON TOTAL-GROUP CORRELATION.

If an investigator is in a situation in which he desires to know what correlation he would obtain if he mingled the data from all groups on a single scatter diagram, and if he knows the sample sizes (n_j), the correlations ($r_{X_jY_j}$), the means, and the standard deviations (σ_{X_j} and σ_{Y_j}) for all the groups ($1, \cdots, j, \cdots, m$), he can compute such a coefficient by means of Formula 1, presented in a slightly different form by Dunlap (1).

$$r_{XY} = \frac{\sum_{j=1}^{m} n_j \sigma_{X_j} \sigma_{Y_j} r_{X_jY_j} + \sum_{j=1}^{m} n_j \delta_j \Delta_j}{\sqrt{\sum_{j=1}^{m} n_j (\sigma^2_{X_j} + \delta^2_j)} \sqrt{\sum_{j=1}^{m} n_j (\sigma^2_{Y_j} + \Delta^2_j)}} \qquad [1]$$

where δ_j is the difference between the mean of the X values for group j and the mean of X for all cases, and Δ_j is the difference between the mean of the Y values for group j and the mean of Y for all the cases. (In this paper it is assumed that all σs are computed using n_j in the denominator. If n_j-1 is used, this value should replace n_j in each formula.)

The above procedure for obtaining an over-all estimate of the degree of correlation between the two variables might be appropriate if the investigator was in a practical situation in which he could not use knowledge of the group from which each subject comes. However, in theoretical studies an investigator is often interested in ascertaining the degree of relationship between the variables if all groups of subjects had the same opportunity or training. Then it is more useful to think in terms of the correlation when the groups have been placed on the scatter diagram so that the means all coincide. To obtain a value for the correlation under this condition, if the sampling of *individuals* has been random and independent from populations that are the same with regard to correlation, one may compute the weighted average correlation coefficient directly or by means of Fisher's z transformation. However, in the usual case where one has used intact groups of subjects the independent random sampling is of *groups* rather than of individuals, and here Lindquist (3, pp. 219–221) recommends that one compute a within-groups correlation by use of analysis of covariance. If the groups do not differ widely in correlation, Lindquist indicates that the resulting coefficient may be treated as though it had been obtained from a simple random sample of

$$\sum_{j=1}^{m} n_j - m + 1$$

individuals, where m is the number of groups. The degrees of freedom would be

$$\sum_{j=1}^{m} n_j - m.$$

Although Lindquist discusses the computation of the within-groups correlation (3, pp. 219 ff.), he leaves the reader who is not well acquainted with analysis of variance and covariance procedures to piece together from scattered parts of his book a computational procedure. McNemar (4, p. 321, p. 327) mentions the within-groups coefficient, but he does not discuss its use. It is mentioned in other scattered references. The present note brings together the discussion of its use and two approaches for computing the within-groups correlation. Which computational approach is to be used depends on the investigator's other manipulations of the data.

If the investigator is not particularly interested in examining the individual correlations within each group of subjects, and is not interested in the means and standard deviations of the variables for his groups, it is efficient to use the raw-score formula, 2, where the individuals in group j are numbered $1, \cdots, i, \cdots, k$, and $k=n_j$.

$$r_{XY} = \frac{\sum_{j=1}^{m}\sum_{i=1}^{k} X_{ij}Y_{ij} - \sum_{j=1}^{m}\left(\frac{\sum_{i=1}^{k} X_{ij}\sum_{i=1}^{k} Y_{ij}}{k}\right)}{\sqrt{\sum_{j=1}^{m}\sum_{i=1}^{k} X^2_{ij} - \sum_{j=1}^{m}\frac{\left(\sum_{i=1}^{k} X_{ij}\right)^2}{k}}\sqrt{\sum_{j=1}^{m}\sum_{i=1}^{k} Y^2_{ij} - \sum_{j=1}^{m}\frac{\left(\sum_{i=1}^{k} Y_{ij}\right)^2}{k}}} \quad [2]$$

In the situation in which the investigator has already computed the correlations and standard deviations within the individual groups of subjects and also wants an over-all evaluation of the degree of relationship between the variables, he may compute the within-groups correlation by means of an equivalent formula, 3, suggested to the writer by S. S. Wilks.

$$r_{XY} = \frac{\sum_{j=1}^{m} n_j r_{X_jY_j} \sigma_{X_j} \sigma_{Y_j}}{\sqrt{\sum_{j=1}^{m} n_j \sigma^2_{X_j}} \sqrt{\sum_{j=1}^{m} n_j \sigma^2_{Y_j}}}. \quad [3]$$

Here $r_{X_jY_j}$ is the correlation between X and Y in group j, σ_{X_j} is the standard deviation of variable X in group j, and σ_{Y_j} is the standard deviation of variable Y in group j.

Before correlations are averaged (directly or by means of the z transformation) they may be corrected for attenuation due to unreliability. In the case of within-groups correlations Wilks has shown the writer that the correction for attenuation may be incorporated directly into the computations. For various groups of subjects reliability estimates of a measure are likely to differ due to different standard deviations for the measure in different groups. In Formulas 4 and 5, corresponding to 2 and 3, respectively, but introducing the correction for attenuation, $r_{X_jX_j}$ and $r_{Y_jY_j}$ are reliability coefficients for variables X and Y, computed separately for each group of subjects. The correction for attenuation is introduced by substituting Formula 3 into the standard formula for correction for attenuation.

$$r_{wcc} = \frac{\sum_{j=1}^{m} \sum_{i=1}^{k} X_{ij} Y_{ij} - \sum_{j=1}^{m} \left(\frac{\sum_{i=1}^{k} X_{ij} \sum_{i=1}^{k} Y_{ij}}{k} \right)}{\sqrt{\sum_{j=1}^{m} \left[r_{X_jX_j} \left\{ \sum_{i=1}^{k} X^2_{ij} - \frac{\left(\sum_{i=1}^{k} X_{ij} \right)^2}{k} \right\} \right]} \sqrt{\sum_{j=1}^{m} \left[r_{Y_jY_j} \left\{ \sum_{i=1}^{k} Y^2_{ij} - \frac{\left(\sum_{i=1}^{k} Y_{ij} \right)^2}{k} \right\} \right]}}. \quad [4]$$

$$r_{wcc} = \frac{\sum_{j=1}^{m} n_j r_{X_jY_j} \sigma_{X_j} \sigma_{Y_j}}{\sqrt{\sum_{j=1}^{m} n_j r_{X_jX_j} \sigma^2_{X_j}} \sqrt{\sum_{j=1}^{m} n_j r_{Y_jY_j} \sigma^2_{Y_j}}}. \quad [5]$$

The assumptions of analysis of covariance are discussed by Jackson (2) who provides illustrations of tests of these assumptions.

REFERENCES

1. DUNLAP, J. W. Combinative properties of correlation coefficients. *J. exp. Educ.*, 1937, **5**, 286–288.
2. JACKSON, R. W. B. *Application of the analysis of variance and co-variance method to educational problems.* Dept. of Educ. Res., Univer. of Toronto, 1940, Bulletin **11**.
3. LINDQUIST, E. F. *Statistical analysis in educational research.* New York: Houghton Mifflin, 1940.
4. MCNEMAR, Q. *Psychological statistics.* New York: Wiley, 1949.

Received April 9, 1956.

PSYCHOLOGICAL BULLETIN
Vol. 54, No. 2, 1957

A GENERAL METHOD OF ANALYSIS OF FREQUENCY DATA FOR MULTIPLE CLASSIFICATION DESIGNS

J. P. SUTCLIFFE
University of Sydney

Fisher (3) has shown the advantages of the factorial experiment over the classical method of "one variable." The following gains accrue from consideration of the effects of independent variables (treatments) upon the dependent variable in the context of other independent variables: (*a*) With a sample of size n, and k treatment classifications which do not interact, "hidden replication" enables estimation of *all* k main effects with the same precision as would be achieved for *one* in a single factor experiment of the same size. The economy of the factorial design is indicated by the fact that to obtain the same amount of information by the "rule of one variable," one would need k sets of n replicates. (*b*) If there is interaction among the treatments, the factorial arrangement enables its isolation and evaluation and thereby sets the limits of generalization. One can specify the effect of the independent upon the dependent variable in a variety of contexts; and conversely, if interaction is zero one may conclude that the relationship is constant through all contexts considered. (*c*) A further virtue of the factorial design lies in the information it may provide about the relative efficacy of different combinations of conditions for the production of given effects. Most use has been made of this in agriculture and industry, but it has its scientific as well as its technological applications, such as in sorting out necessary and sufficient conditions.

In practice the factorial design has most often been used where it is possible to obtain *measurements* on the dependent variable, so that statistical analysis of the outcomes is by way of the "analysis of variance." In many research areas, however, phenomena are not yet amenable to scaling so that one has counts or frequencies within given categories rather than measures, e.g., male versus female rather than degrees of sexuality. There is no logical hindrance to the use of factorial experimentation with these phenomena, and such is to be recommended in light of the advantages to be gained. The problem is to find a method of statistical analysis appropriate to this type of data. χ^2 methods are available for the comparison of sampled frequencies and for assessing association in simple contingency tables. These cases are in effect instances of single and double classification designs, and if contingency association is the analogue of interaction in analysis of variance, then a method of assessing multiple contingency is needed for the analysis of frequency data from higher order designs. Pearson (6) described a procedure for assessing multiple contingency but failed to consider the question of additivity of χ^2 components. Bartlett (1) offered a method for the $2\times2\times2$ case which involves the solution of a cubic equation and is difficult to apply in practice. Recently, Lancaster (5) following proofs by Irwin (2) and Lancaster (4) has devised a general method of partitioning a total χ^2 and degrees of freedom into independent additive components due to given sources of variation. This completes

the parallel with the analysis of variance in which the total sum of squares and degrees of freedom are partitioned into sums of squares and *df* for all main effects, interactions and error. This paper presents for psychologists a general form of multiple contingency analysis based on Lancaster's work, provides an illustration, and comments on the generality of application of the method.

Multiple Contingency Analysis

Complex contingency tables of frequency data from multiple classification designs may be of several forms according as the sampling of main effects is random or restricted. (*a*) The *random* case imposes no sampling restrictions. For example, after a random sample has been taken, it may be classified in various ways and the frequencies within classes will be due only (within sampling limits) to the population proportions. (*b*) The *mixed* case involves restrictions upon the proportions within categories of given classifications and freedom with respect to others, e.g., arranging in advance that a total sample will involve equal proportions of the sexes. Parameters for a classification are defined by its restriction. Whichever case is involved, for each *observed* frequency in the table there will be an *expected* value, and hence divergence of the total table from expectation may be tested through χ^2. Within a total table, however, there will be a number of sources of variation comprising main effects and interactions, and to isolate them one would need to partition the total χ^2 and *df* into independent additive components due to such sources. The problem is to specify the expected frequencies which will meet this requirement. The method will be developed through a notation which, while perfectly general, will for simplicity be set out for an $A \times B \times C$ design.

Let the classifications be symbolized as $A, B, C, \cdots, L$. Let A be subdivided into a categories represented generally by the subscript i which thus takes values from 1 to a. Similarly B is represented by $(j = 1, \cdots, b)$, C by $(k = 1, \cdots, c)$, etc. Let p_{ijk} = the probability of an observation falling in the ijkth cell; o_{ijk} = the observed frequency in the ijkth cell; and e_{ijk} = the expected frequency in the ijkth cell. Let a dot in place of a subscript represent summation across the values represented by the subscript, e.g.,

$$\sum_{i=1}^{i=a} o_{ijk} = o_{.jk}.$$

Let the total sample size $o_{...} = N$; and finally $p_{...} = 1.0$. On the hypothesis[1] of zero interaction, $p_{ijk} = p_{i..} \times p_{.j.} \times p_{..k}$, $p_{ij.} = p_{i..} \times p_{.j.}$, etc. These parameters are used to find the expected frequencies, e.g., $e_{ijk} = p_{ijk} \times N$, and hence χ^2 may be calculated as $\sum(o-e)^2/e$. Now some or all of the values of the parameters $p_{i...}$, $p_{.j.}$, $p_{..k}$ may be (*a*) known from the population; or (*b*) estimated from the sample, e.g., $p_{i..} = o_{i..}/N$. These situations taken with the random and mixed designs provide four cases each requiring separate consideration. Case (*1a*) will be presented in full for the $A \times B \times C$ design.

(*1a*) *Random sampling, known parameters*

The partition of total χ^2 and *df* into component values for this case,

[1] Ordinarily one works with the null hypothesis, but population hypotheses of non-zero interactions may be entertained, e.g., in a test of goodness of fit with case *1a*, and again in determining the power of the test of significance for a given situation.

TABLE 1

PARTITION OF χ^2 AND *DF* FOR CASE (1*a*)

Number	Source	χ^2	*df*
1	*A*	$\chi^2_A = \sum^{a} (o_{i..} - e_{i..})^2/e_{i..}$	$(a-1)$
2	*B*	$\chi^2_B = \sum^{b} (o_{.j.} - e_{.j.})^2/e_{.j.}$	$(b-1)$
3	*C*	$\chi^2_C = \sum^{c} (o_{..k} - e_{..k})^2/e_{..k}$	$(c-1)$
4	*AB*	$\chi^2_{AB} = \sum^{a}\sum^{b} (o_{ij.} - e_{ij.})^2/e_{ij.} - (1+2)$	$(a-1)(b-1)$
5	*AC*	$\chi^2_{AC} = \sum^{a}\sum^{c} (o_{i.k} - e_{i.k})^2/e_{i.k} - (1+3)$	$(a-1)(c-1)$
6	*BC*	$\chi^2_{BC} = \sum^{b}\sum^{c} (o_{.jk} - e_{.jk})^2/e_{.jk} - (2+3)$	$(b-1)(c-1)$
7	*ABC*	$\chi^2_{ABC} = \sum^{a}\sum^{b}\sum^{c} (o_{ijk} - e_{ijk})^2/e_{ijk} - (1+2+3+4+5+6)$	$(a-1)(b-1)(c-1)$
8	Total	$\chi^2_T = \sum^{a}\sum^{b}\sum^{c} (o_{ijk} - e_{ijk})^2/e_{ijk}$	$(abc-1)$

together with computing formulas, are set out in Table 1. As the population values are known, the significance of all main effects and interactions may be assessed.

(*1b*) *Random sampling, parameters estimated from the data*

In this case one estimates population proportions from the sample data, e.g., $p_{i..} = o_{i..}/N$, and as $e_{i..} = p_{i..} \times N$, then $e_{i..} = o_{i..}$. Accordingly for this case the values of χ^2 and *df* for the main effects are zero. One may assess all interactions and their *df* are unchanged, but the total *df* is reduced by the number lost with the main effects.

(*2a*) *Mixed case, known parameters*

Here restriction specifies the parameters for a classification, so that the main effects and *df* for the restricted classifications are zero. Furthermore, if several classifications and their subclasses are restricted, *within* that set the interactions and *df* are also zero. For instance, if in an $A \times B \times C$ design the proportions within the A and B classes and AB classes are prearranged and sampling is random only with respect of C, then one has set $e_{i..} = o_{i..}$, $e_{.j.} = o_{.j.}$, $e_{ij.} = o_{ij.} = (o_{i..} \times o_{.j.})/N$. In this case one would obtain

$$\chi^2_T = \chi^2_C + \chi^2_{AC} + \chi^2_{BC} + \chi^2_{ABC},$$

and the total *df* would be reduced by the number lost with the main effects and interactions.

(*2b*) *Mixed case, parameters estimated from the data*

Here one loses all the main effects and such interactions as involve only the restricted classifications. For the case with A and B restricted and C random,

$$\chi^2_T = \chi^2_{AC} + \chi^2_{BC} + \chi^2_{ABC},$$

and as before the total *df* has to be adjusted for the number lost with the main effects and interactions.

ILLUSTRATION OF THE METHOD

An experiment is reported (**7**) in which the manner of resolving conflict (A, $i = 1, 2$) is observed under four conditions constituting the fac-

torial arrangement of two conditions of social distance (B, $j=1, 2$) and two conditions of publicity (C, $k=1, 2$). Four independent random samples of 100 cases were assigned to the four conditions, and the whole experiment was replicated for eighteen conditions of social sanction (D, $l=1, 2, \cdots, 18$). In this way equal numbers were subjected to all treatments and the only main effect frequencies free to vary were those pertaining to type of conflict resolution. That is, A is random and B, C, and D and their subclasses are restricted in an $A\times B\times C\times D$ design. As the population proportions for type of conflict resolution were unknown, they were estimated from the sample data. Hence the analysis follows the (*2b*) type, where

$$\chi^2_T=\chi^2_{AB}+\chi^2_{AC}+\chi^2_{AD}+\chi^2_{ABC}+\chi^2_{ACD}+\chi^2_{ABD}+\chi^2_{ABCD}$$

$$df_T=(bcd-1)(a-1).$$

Generality of Application

As presented the method has application to factorial experiments in which information on the dependent variable is in frequency form. Equally the method may be applied to surveys where sampling units are classified in a variety of ways. A further application of this type of method to measurement data has recently been suggested by Wilson[2] (**8**) who uses it as a "distribution-free" substitute for the analysis of variance. He does not justify this substitution and some comment is warranted. Wilson dichotomizes his measures at the median and, in effect, introduces the dependent variable as an additional classification with two levels. Information is lost in categorizing and to that extent a test of significance with frequencies is less sensitive than one applied to measures. Hence one would only use with measurement data multiple contingency analysis as a substitute for analysis of variance when the latter method was not applicable. This would be so when certain assumptions required for a valid F test could not be met—normality of parent population, homogeneity of variance—and a suitable transformation was not available. Here in the absence of the more sensitive test, the less sensitive test would certainly be preferable to none at all.

[2] Wilson's procedures are based upon a particular hypothesis about the expected values, viz., irrespective of treatment effects a score on the dependent variable is equally likely to occur above or below the median. It should be noted that this is not the only population hypothesis which may be entertained. The method presented in this paper, being more general than Wilson's, is to be preferred on that score.

REFERENCES

1. Bartlett, M. S. Contingency table interactions. *J. roy. statist. Soc. Suppl.* 1935, **2**, 248–252.
2. Irwin, J. O. A note on the subdivision of χ^2 into components. *Biometrika*, 1949, **36**, 130–134.
3. Fisher, R. A. *The design of experiments.* (5th Ed.) Edinburgh: Oliver & Boyd, 1949.
4. Lancaster, H. O. The derivation and partition of χ^2 in certain discrete distributions. *Biometrika*, 1949, **36**, 117–129.
5. Lancaster, H. O. Complex contingency tables treated by the partition of χ^2. *J. roy. statist. Soc.*, Series B, 1951, **13**, 242–249.
6. Pearson, K. On the theory of multiple contingency with special reference to partial contingency. *Biometrika*, 1915–17, **11**, 145–158.
7. Sutcliffe, J. P., & Haberman, M. Factors influencing choice in role conflict situations. *Amer. social. Rev.*, December, 1956, **21**.
8. Wilson, K. V. A distribution-free test of analysis of variance hypotheses. *Psychol. Bull.* 1956, **53**, 96–101.

Received May 2, 1956.

PSYCHOLOGICAL BULLETIN
Vol. 54, No. 2, 1957

AN AID IN THE COMPUTATION OF CORRELATIONS BASED ON Q SORTS

JACOB COHEN[1]

Franklin D. Roosevelt VA Hospital, Montrose, New York, and New York University

With the increase in recent interest in Q technique stimulated primarily by the work of Cattell (1) and Stephenson (3), researchers in the area of clinical-personality-social have needed frequently to compute Pearsonian product-moment coefficients of correlation from Q arrays in large numbers. Many of them have simply followed the standard textbook computing methods which, because of the special conditions involved in correlating Q arrays, are exceedingly inefficient. The purpose of this note is to present a quick method for computing such correlations.

The special conditions referred to above are these: All the Q arrays to be intercorrelated have, because of the instructions for Q sorting, exactly the same distribution, and therefore identical means and standard deviations. The most efficient formula for computing rs under these circumstances is derived from the formula for the product-moment r by the "method of differences" (2, p. 118, formula 186).

$$r_{12}=\frac{\sigma^2{}_1+\sigma^2{}_2-\sigma^2{}_D}{2\sigma_1\sigma_2}, \qquad [1]$$

where D represents the difference between paired scores.

Since the two distribution standard deviations are equal, this formula reduces to

$$r=\frac{2\sigma^2-\sigma^2{}_D}{2\sigma^2}=1-\frac{\sigma^2{}_D}{2\sigma^2}. \qquad [2]$$

[1] From the Psychology Service, Franklin D. Roosevelt Veterans Administration Hospital, Montrose, New York.

The $\sigma^2{}_D$ term can be readily expressed in raw-score terms as follows:

$$\sigma^2{}_D=\frac{\sum D^2}{N}-M^2{}_D, \qquad [3]$$

where M_D represents the mean of the paired differences.

Since the mean of the paired differences must equal the difference between the means, which is zero,

$$\sigma^2{}_D=\frac{\sum D^2}{N}. \qquad [4]$$

Substituting in Equation 2,

$$r=1-\frac{\sum D^2}{2N\sigma^2}. \qquad [5]$$

In any given Q-technique research, the denominator of the fraction is a constant, K, for all the correlations to be performed, since both N, the number of statements, and σ^2, the variance of the forced frequency distribution of scale values, are constant. Equation 5 therefore becomes

$$r=1-\frac{\sum D^2}{K}. \qquad [6]$$

For any given correlation, the $\sum D^2$ is readily found and substituted for the arithmetic computation of r.

Since Equation 6 is a linear equation, however, whenever the number of correlations to be found becomes large, it is a simple matter to construct a nomograph and read off the values of r (see Fig. 1). The nomograph is constructed as follows:

1. On a sheet of graph paper ori-

ented so that the long sides are at left and right, lay off along the left side values of $\sum D^2$ from zero at the bottom to K at the top, and on the right values for $\sum D^2$ from K at the bottom to $2K$ at the top.

2. On the bottom horizontal lay off values of r from $+1.00$ at the left to .00 at the right. On the top horizontal lay off values of r from .00 at the left to -1.00 at the right.

3. Draw a straight line from the lower left corner ($r = 1.00$, $\sum D^2 = 0$) to the upper right corner ($r = -1.00$, $\sum D^2 = 2K$).

In using the nomograph, when $\sum D^2$ is entered from the left, r is read off at the bottom; when $\sum D^2$ is entered from the right, r is read off at the top.

Numerical illustration. A set of 100 statements is Q sorted into a forced distribution with a σ^2 on the Q scale of 4.0. K therefore equals $2(100)(4.0) = 800$ and Equation 6 becomes

$$r = 1 - \frac{\sum D^2}{800}.$$

For illustrative purposes, the nomograph of this equation is given in Fig. 1.

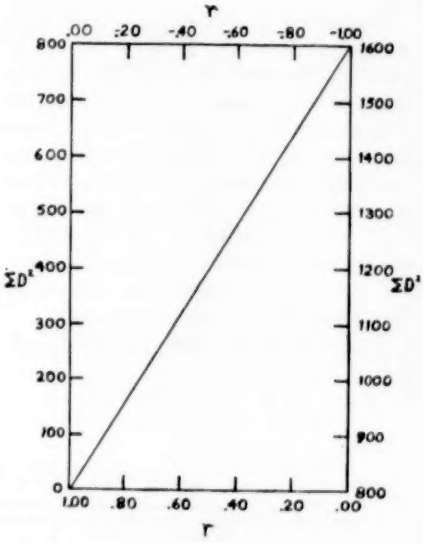

FIG. 1. EXAMPLE OF NOMOGRAPH WHEN $K = 800$.

REFERENCES

1. CATTELL, R. B. *Factor analysis.* New York: Harper, 1952.
2. DUNLAP, J. W., & KURTZ, A. K. *Handbook of statistical nomographs, tables, and formulas.* Yonkers, N. Y.: World Book Co., 1932.
3. STEPHENSON, W. *Study of behavior: Q-technique and its methodology.* Chicago: Univer. of Chicago Press, 1953.

Received May 7, 1956.

PSYCHOLOGICAL BULLETIN
Vol. 54, No. 2, 1957

GRAPHIC DETERMINATION OF SIGNIFICANCE OF 2×2 CONTINGENCY TABLES

DAVID K. TRITES
School of Aviation Medicine, USAF, Randolph Field, Texas

A number of techniques have been published (e.g., **1**, **2**, **3**, **4**, **5**, **6**) to assist in the evaluation of the significance of differences in proportions, or frequencies, arranged in 2×2 contingency tables. All of the available methods suffer from limitations either in generality of problems to which they may be applied, in the arithmetic computations involved in their use, or in the appropriate nature of the obtained probability levels.

The present method was designed to overcome two of these limitations. The arithmetic operations have been reduced to simple addition and subtraction of the cell entries and the obtained significance levels are based on the chi-square distribution. The one limitation of the technique is the requirement that for maximum usefulness the two samples being compared must be independently selected and contain the same number of subjects. Nevertheless, it is still possible in certain cases where the require-

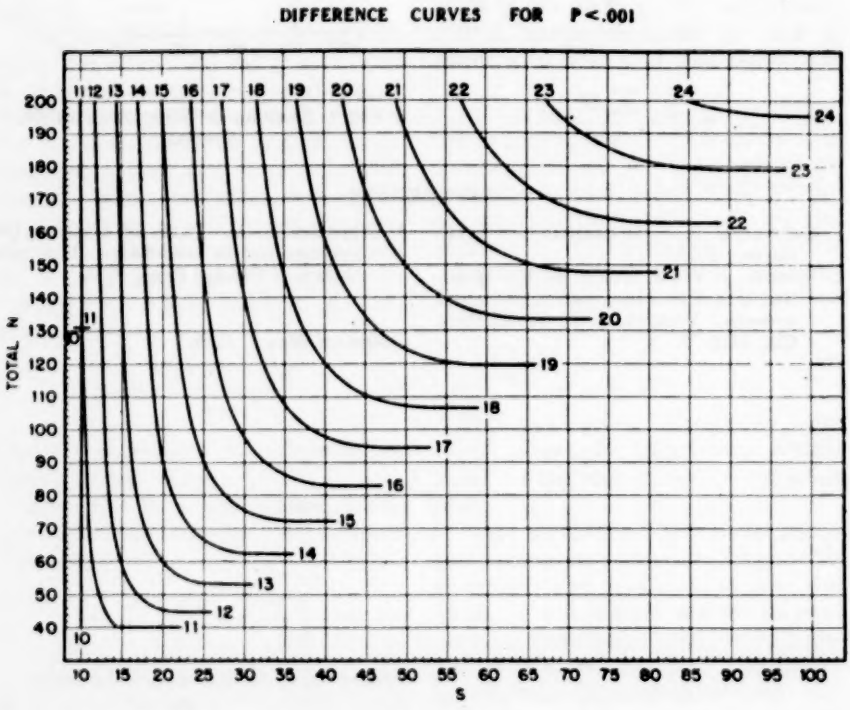

FIG. 1

ment of equal sample frequencies is not met to obtain a conservative probability statement.

In developing the present technique, use was made of certain relationships inherent in the computing formula for chi square for 2×2 contingency tables. Consider the following table.

	Sample 1	Sample 2	
Classification X	a (21)	b (9)	$a+b=(30)$
Classification Y	c (24)	d (36)	$c+d=(60)$
	$a+c$ (45)	$b+d$ (45)	$a+b+c+d=N=$ (90)

where a, b, c, and d are cell frequencies.

If the two column totals, $a+c$ and $b+d$, are equal and if $a+b$ is chosen to be equal to or less than one-half the total number of subjects ($\frac{1}{2}N$),[1] and further if we set

$$a+b=S \quad \text{and} \quad a-b=D,$$

it can be shown that

$$\chi^2=\frac{D^2N}{S(N-S)} . \qquad [1]$$

Solving this for N gives

$$N=\frac{S^2\chi^2}{\chi^2S-D^2} . \qquad [2]$$

[1] The restriction that the smaller of the two row totals be used is not necessary for the development of Formula 1. However, in order to increase the legibility of the charts, the requirement was imposed.

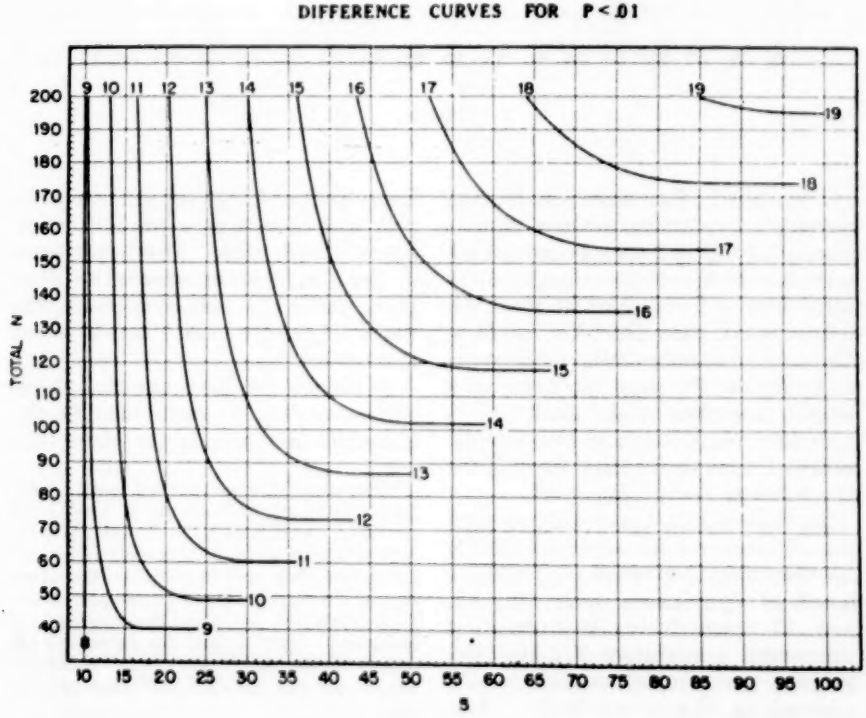

FIG. 2

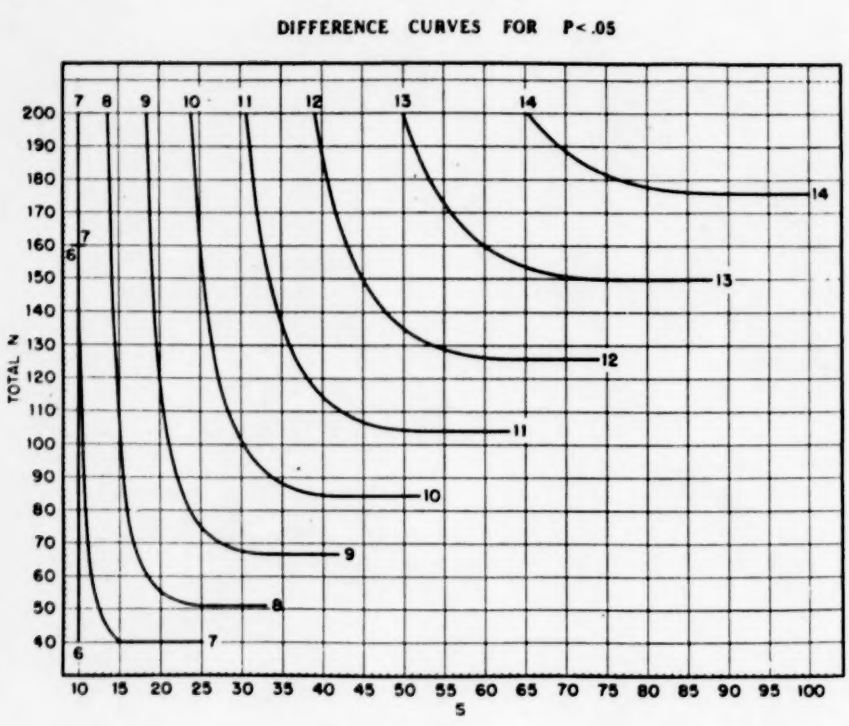

FIG. 3

For a fixed value of chi square, representing any desired level of significance, Equation 2 may be solved for all values of N corresponding to varying values of S and D which meet the requirement that $S \leqq \frac{1}{2}N$. If D is also fixed, and only S allowed to vary, a series of Ns may be computed which, together with their corresponding Ss, determine the coordinates of a curve for each fixed D for the selected probability level.

In Figures 1, 2, 3, and 4 are plotted the N, S curves for several fixed Ds for the critical values of $\chi^2(df = 1)$ for levels of significance .001, .01, .05, and .10, respectively. In computing the curves, a maximum N of 200 (100 subjects per sample) was arbitrarily selected as the upper limit. The lower limit of 40 (20 subjects per sample) was chosen as a prudent minimum sample size. The lower limit of the S values was selected as 10 so that, under the null hypothesis, the value in cells a and b would be at least five.[2]

It should also be noted that the D value assigned to each curve in the charts is one unit larger than that

[2] The chi-square values used were: 10.827 for the .001 level; 6.635 for the .01 level; 3.841 for the .05 level; and 2.706 for the .10 level. Computations were done twice, independently, and then spot checked a third time. During the plotting of each curve any erratic fluctuations were noted and the computations rechecked. It is felt that the curves are accurate within the limits of the numerical values used and the usual limitations of plotting.

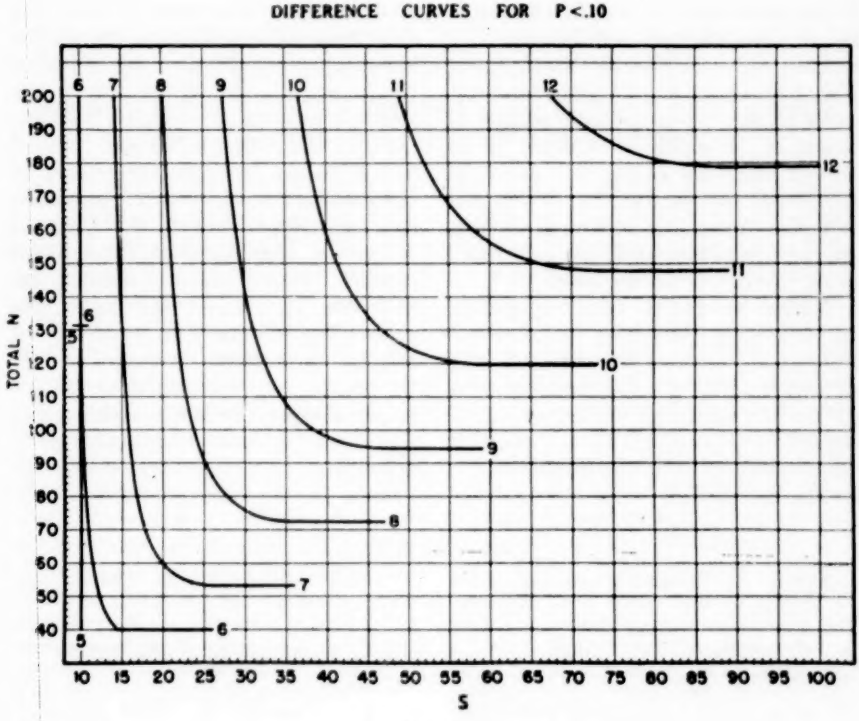

FIG. 4

used in the computation of that curve. For example, using a chi-square value of 2.706 ($p=.10$) the curve obtained for a D value of 7 would be labeled on the chart as 8; that for a D value of 8 labeled as 9; and so on. This is necessary since the region between any two curves (referring now to the D values actually used in the calculations), for example $D=7$ and $D=8$, represents values greater than 7 but less than 8. Since in practice only integral D values may be obtained, all the fractional values greater than 7 are theoretical only. Actually a difference of 8 must be obtained before significance can be reached. Consequently, the $D=7$ curve must be labeled $D=8$.

To use the charts, it is first necessary to determine which pair of cell entries, $a+b$ or $c+d$, is the smaller. Choose $S=$smaller $(a+b$ or $c+d)$. With this sum (S) on the abscissa and the total number of cases in both samples (N) on the ordinate enter the chart appropriate for the desired level of significance. The difference curve (D) immediately to the left of the intersection of the S and N values is the minimum difference between the two selected cell frequencies required for statistical significance.

As an example, consider the numerical values given in parentheses in the preceding 2×2 table. The two row totals are 30 and 60. The chart is entered with $S=30$ (since $30<\frac{1}{2}N=45$) and with $N=90$. To determine

if the difference of $D=21-9=12$ is larger than what one would expect by chance at, say, the .01 level, enter Fig. 2 and note the D curve immediately to the left of the intersection of S and N. For this example, it is observed that any $D \geqq 12$ will result in statistical significance for $S=30$ and $N=90$. Since our observed $D=12$, we may conclude that the difference is significant at the .01 level.

Yates' correction may be applied by subtracting one from the obtained difference between the selected pair of cell values. In the example just given, the obtained difference of 12 would be reduced to 11 and the difference would no longer be significant at the .01 level.

When the frequencies in the two samples are unequal, a test of significance may be obtained by reducing the larger sample to the size of the smaller, keeping the original proportions in the cells unchanged. The S and D values are then determined in the usual manner and the appropriate chart entered with a total N equal to twice the frequency of the smaller sample. If the graph indicates significance for this comparison, the probability of the difference in the original data must be even smaller. Of course, if after reduction of the larger sample the comparison is not significant, no statement can be made regarding the significance of the difference in the unreduced data.

REFERENCES

1. APPEL, V. Companion nomographs for testing the significance of the differences between uncorrelated proportions. *Psychometrika*, 1952, **17**, 325–330.
2. LAWSHE, C. H., & BAKER, P. C. Three aids in the evaluation of the significance of the difference between percentages. *Educ. psychol. Measmt*, 1950, **10**, 263–270.
3. MAINLAND, D., & MURRAY, I. M. Tables for use in fourfold contingency tests. *Science*, 1952, **116**, 591–594.
4. PETERSEN, R. L. *A graphic method for estimating the significance of differences between proportions or percentages.* Washington, D. C.: ARDC, Bolling AFB, 1954. (Hum. Factors Operat. Res. Lab., Mem. No. TN-54-6).
5. SELLS, S. B., FRESE, F. J., & LANCASTER, W. H. Research on the psychiatric selection of flying personnel: II. Progress on development of SAM Group Ink-Blot Test, Appendix VI. *USAF Sch. Aviat. Med. Proj. Rep.*, 1952, Proj. No. 21-37-002 (Rep. No. 2).
6. ZUBIN, J. Nomographs for determining the significance of the differences between the frequencies of events in two contrasted series or groups. *J. Amer. statist. Ass.*, 1939, **34**, 539–544.

Received May 21, 1956.

PSYCHOLOGICAL BULLETIN
Vol. 54, No. 2, 1957

THE USE OF THE SPLIT-LITTER TECHNIQUE IN PSYCHOLOGICAL RESEARCH[1]

SHERMAN ROSS, BENSON E. GINSBURG, AND VICTOR H. DENENBERG
The University of Maryland, The University of Chicago, and Purdue University

Psychologists, as careful scientists, have in their research and teaching been greatly concerned with the problems of methods and experimental controls. One of the common methods used in animal experimentation is the litter-mate control or split-litter technique. Munn in his discussion of general procedures in research with the rat presents the rationale underlying the use of this technique when he states, "In many experiments it is necessary to use two or more groups of comparable genetic constitution, a control group and one or more experimental groups. The closest one can come to achieving comparability is to have the different groups consist of litter mates" (**8**, p. 7).

Munn refers to a paper by Corey (**2**) which was concerned with the problem of the initial equating of control and experimental groups. In 1930 Corey pointed out the " . . . misplaced confidence in the uniformity of the subjects that is supposed to be gained through the use of inbred stock" (**2**, p. 287). His evidence was gained on 160 rats from a colony bred for about one year from six pairs of Wistar rats. Coefficients of correlation for learning performance between litter halves were: trials, .78; active time .72; errors, .80; and total time, .30. It is clear from these data that there is a pronounced correlation between litter halves indicating significant litter differences. Further there is the implication that differences in initial (genetic) ability might be the basis for these litter differences. Corey suggested as an experimental design the selection of animals to represent a normal distribution of the ability of all *S*s. He admits that this is difficult and suggests the usefulness of the split-litter technique, pointing out the obvious value of this technique in serving to hold constant certain environmental factors. Corey, however, does not specifically deal with the problem of genetic control by means of the split-litter method.

Experimental evidence such as we have just cited, in addition to the considered opinions of many psychologists, tends to lead to the acceptance and wide use of a given technique. Quite often certain fundamental assumptions are accepted without critical appraisal. We are concerned here with a tacit assumption which we believe has been accepted by many psychologists. The assumption is that when the split-litter technique is used, all genetic factors pertinent to the variable under investigation are held constant. Thus, any differences obtained are strictly a function of environmental factors, presumably the treatment effects introduced into the experimental situation. The assumption is expressed or implied in the following statements made by highly competent research workers. "But with such small groups, and especially without litter-mate controls, this

[1] This paper was prepared at the Division of Behavior Studies, R. B. Jackson Memorial Laboratory, Bar Harbor, Maine, during the summer of 1955. V.H.D. was a Carnegie Fellow, S.R. and B.E.G. are Scientific Associates of the Laboratory.

conclusion is entirely gratuitous" (8, p. 347). "The only control over possible genetic differences was the division of each litter into solitary animals and normal animals" (1, p. 73). " . . . When heredity is held constant between experimental groups, by the split-litter method . . . " (5, p. 533).

The geneticist, however, looks at this problem differently than does the psychologist. Scott (10) and more recently Ginsburg (3) have pointed to the importance of the genetic characteristics of the animal *Ss* used by psychologists.

Scott discusses the relationships between genetic variables and behavioral variability. In regard to the use of litter-mate controls he says, " . . . litter mates show on the average a correlation of .5 in hereditary variables. However, since an animal gets half of its variable heredity from each parent it is theoretically possible in small samples to get litter mates which are entirely different. Litter-mate controls should be considered as a means of controlling age and environment, although it has been shown that since the animals affect each other, and since the eggs may be lodged in different parts of the uterus, the environment is far from identical for each animal" (10, p. 529).

Ginsburg in discussing the problem of the genetic variables which are frequently ignored in psychological investigations says, "In a small or moderate sized animal colony that has not been rigorously inbred or subjected to the strictest selection with respect to the trait in question, the use of litter-mates does not constitute an adequate genetic control. If there is appreciable heterozygosity for genetic factors affecting the experimental outcome, these will segregate among litter-mates, making them genetically unlike each other. Under the conditions just discussed, it would be preferable to select experimental and control samples at random from the colony, than to use the split-litter technique where, whatever the number of litters or animals involved in the latter procedure, they trace back to a limited number of relatively recent matings within the colony" (3, p. 41).

We believe that Ginsburg's definition of the type of animal colony for which the split-litter technique is ***not*** an adequate genetic control is descriptive of many of the animal colonies which are used by psychologists as sources of experimental *Ss* (probably including Corey's *Ss*). First, the colonies are small to moderate in size (*N* less than 500). Second, the strains used are rarely pure strains. In fact most departments which maintain rat colonies for experimental purposes have only partially inbred lines. These animals may differ markedly one from the other, even though started from a (roughly) genetically similar pair of rats. Since these are partially inbred strains, and strains in which the selection for the ***traits under experimental investigation*** is relatively unknown, there may be segregation of genes which can affect the outcome of the experiment.

A consideration of genetic theory will easily demonstrate why this is so. In order to achieve complete genetic uniformity, we must produce individuals whose genotypes are identical. In so far as they are different, we may expect to get variability which is compounded due to the recombination of genes at each mating, much as cards are recombined after being shuffled and dealt anew. The phenotypic value of a given gene depends, in part, upon the genes with which it interacts, just as the value to the card player of the queen of hearts depends

in part on the other cards he holds in his hand. The magnitude of phenotypic variability, which is the factor of greatest interest to the psychological researcher, is a function of the possible combinations of genes rather than of their actual number. Inbreeding reduces the field of genetic variability but may actually *increase* phenotypic variability in some cases (9). These include situations where a given genotype that may become fixed as a result of inbreeding is more susceptible to environmental influences than most other genotypes, and cases of multiple factors affecting a quantitatively varying character. In the latter case, if homozygosity represents the extreme phenotypic types, partial inbreeding will increase the standard deviation of the population with respect to the character as the homozygous types increase at the expense of the heterozygotes. Calculations of the degree of genetic relationship probably achieved in application by a given system of inbreeding over a known number of generations may be easily made (11, 12). While such calculations of genetic relationship will yield valuable information regarding the expected decrease in heterozygosis under a given system of inbreeding, this information is with respect to either the average heterozygosity for all genes in a given animal or the average condition for a pair of alleles in the population as a whole. This is not particularly useful when we are concerned with the possible effects of one or a few specific loci in an experimental situation where all of the hazards of sampling error apply.

The process of mating close relatives, therefore, is, by itself, no guarantee of actual genetic uniformity. Selection, conscious or unscious, may favor heterozygosity for factors affecting fertility, viability, etc. Thus a number of loci may be prevented from stabilizing. There is good evidence that sublines of a single inbred line separated after original inbreeding of more than 20 to 30 generations of brother × sister matings have, after their separation, shown evidence of genetic differences. The basis of these differences is either different segregation of residual heterozygosity in the common parents of both sublines or newly arising mutations. It is difficult to establish which process occurred in any particular case.[2]

The inbred animals, to be sure, are alike for a great many genes, but the genes for which they remain variable would have to be ruled out as affecting the experimental situation before the use of the split-litter method would constitute an adequate control for genetic factors. A test of this is to obtain the parent-offspring correlation for a number of litters on the trait which is of experimental interest. If the correlation is high, two things are indicated. First, genetic factors and/or some environmental factors (which can be experimentally eliminated) are affecting the characteristic being measured. Second, these factors are relatively constant within a litter. In this situation the split-litter method is of value. If the parent-offspring correlation is low, this might be due to any of several reasons, some of which are: (*a*) genetic factors do not affect the trait being measured, (*b*) all genetic factors influencing this trait are identical for all litters, (*c*) recessive genes are segregating, or (*d*) mutations possibly are occurring. Regardless of the basis of the low correlation, splitting litters in this situation is of no value and may even be detrimental to the efficiency of the experiment.

[2] Dr. Elizabeth S. Russell, personal communication, January 12, 1956.

Even highly inbred strains may not remain uniform forever. Thus, the DBA/1 Jax mice, which two of us have used as *S*s, and which constitute as pure an inbred population as one is likely to get, is composed of a number of sublines which have changed from each other in time after having been inbred to the point of practical genetic identity. This was made evident when the population from which we had drawn our *S*s was decimated by the fire at the R. B. Jackson Memorial Laboratory in 1947. DBA/1 mice, descended some generations back from the same ancestors as the ones we had been using, were sent in to the Laboratory to replace the ones destroyed. Some of these animals reacted like the prefire mice, but others did not in a situation where genetic variables were of primary importance (3). The continuing genetic identity of an inbred strain can, therefore, not be taken for granted. Responsible geneticists and laboratories engaged in the production of inbred strains must and do continually check their materials to make sure that the strains remain constant.

We do not mean to imply by this that the method of litter-mate controls should not be used. What we do wish to state is that this procedure in no way guarantees that the genetic factors, which are liable to influence the results of the study, are necessarily held constant when either partially inbred strains or randomly bred animals are used.

Perhaps this point can be further clarified by examining the extreme conditions. If isogenic strains (all *S*s in the colony have identical genes) are used, splitting litters cannot control for hereditary variables since these are already constant. Splitting the litter will in part control for certain environmental factors which are constant within a litter, but differ from litter to litter (e.g., maternal influence and age). At the other end of the continuum, one could use offspring from mongrel animals purchased from pet shops throughout the country and randomly mated. One would not expect the offspring of the animals within a litter to be genetically more similar in regard to the behavior being investigated than offspring from different litters, since all the parents have a random assortment of genes. Here again the split-litter method would equate for certain environmental factors. The *S*s generally used in psychological research are somewhere between these two extremes, possibly nearer the isogenic situation. Thus, we would expect that when partially inbred animals are used, splitting a litter will equate for *some* of the hereditary factors influencing the behavior being studied, but not all of these factors. Genes not yet stabilized within a colony will be segregating within litters and producing genetically unlike litter mates. Here again the splitting will control for certain environmental factors, but these only so far as they are constant within a given litter.

From the statistical point of view the above discussion would seem to indicate that the split-litter method can be of value when used with the appropriate analytical procedures. That is, the split-litter method may be considered to be a stratified random sample in which the litters are the strata and the *S*s within the litters are the sampling units. Thus, if one used a "matched groups" design and removed "litters" as a source of variance, this would appear to remove some (though not all) of the genetic variability influencing the

behavior being measured. The error term would be reduced and greater precision to the test of significance of the independent variable would be achieved. This argument is not necessarily true. Hansen, Hurwitz, and Madow (**4**, ch. 5) in their discussion of stratified sampling point out that biased estimates and loss of precision can occur by sampling from very small strata (litters) and by the use of relatively few cases within each stratum. Under these conditions random sampling is preferable.

A comparison of random and stratified sampling can be made by considering a single locus at which a dominant gene and (in the simplest case) one recessive allele are involved and in which the occurrence of heterozygosity will affect the experimental results. Since, in most cases, we are unable to detect the heterozygous individuals in any obvious way, our problem becomes that of achieving a nearly equal distribution of heterozygotes between our experimental and control groups. To this end we have a choice between sampling the animal colony at random and equating the experimental and control *S*s on the basis of age, sex, weight, etc., or equating the groups by the split-litter technique. Essentially this is a choice of taking either the ***individual*** or the ***mating*** as our unit of sampling. In the former case, the probability of including some heterozygotes in the experimental or control group is a function of the frequency with which such individuals occur in the colony, and of the sampling error. In the latter case, the probability is a function of the frequency with which matings capable of producing such individuals occur in the colony, and again, of the sampling error. On the simplified assumptions of random mating in the absence of mutation and selection affecting the trait in question, and on the further assumption that the genes under consideration are equally distributed in the two sexes, the relative frequency with which a particular genotype may be expected to occur can be calculated either from a consideration of the types of matings possible and the relative frequency of each, or of the gametes available in the colony and the frequency of the genotypes to be expected as a result of their random combinations.

Using the first method, there are six possible matings to consider:

1. AA×AA
2. AA×Aa
3. AA×aa
4. Aa×aa
5. Aa×Aa
6. aa×aa

Of these, No. 6 may be eliminated, since it can be identified by the phenotype and does not contribute to the heterozygosity in any case. Nos. 3 and 4 may also be excluded, even though they do contribute to the heterozygosity of the succeeding generation because they, too, are identifiable phenotypically. No. 5 may be identified through the occurrence of recessive progeny, thus leaving 1 and 2 as the source of our population. On the assumptions enumerated, if X represents the proportion of AA individuals in the parental generation, and Y represents the proportion of Aa individuals, then the frequency with which type 2 ***matings*** (producing heterozygotes) occur may be derived from the expression $(X\ \mathrm{AA} + Y\ \mathrm{Aa})^2$ and is equal to $2XY$. Only half the progeny of such matings, or XY, will consist of heterozygous ***individuals***. It is thus evident that whatever the absolute numbers may be in a given case, when the ***individual***

is taken as the sampling unit, heterozygosis will occur only half as frequently as when the *mating* is taken as the unit. Given the usual circumstances of relatively small numbers of *S*s in each of the groups, the sampling errors become appreciable and the most reasonable supposition is that heterozygosis encountered in a litter that is split between experimentals and controls will not be equally distributed.

Using the second method and the same assumptions, the relative frequency with which a particular genotype may be expected to occur can be obtained by expanding the binomial $[q+(1-q)]^2$ where q equals the frequency of gene A and $(1-q)$ equals the frequency of gene a (**6**, ch. **1**; **7**, ch. 6). These methods may be extended to additional independent pairs of alleles by making the calculation for each pair separately and combining these through the use of the product law. In the case of a multiple allelic series, the gene frequency notation is extended by using a polynomial corresponding to the number and frequency of the alleles involved, rather than a binomial, as in the simpler model cited here. Where the genes are not equally distributed between the sexes, separate polynomials representing the distribution for each sex must be used.

In addition to the statistical and genetic arguments there are several experimental points to consider as well. Though certain of the environmental factors within a litter are constant for all *S*s, there are other factors which are variable and which make for dissimilarity among litter mates. Some of these factors are litter size, competition for food, patterns of dominance and aggression, and differences in uterine environment. If these variable factors are ones which will influence the dependent variable, then splitting litters in no way equates for them. Indeed, this procedure is likely to lead to larger intralitter differences than interlitter differences. Thus the precision of the test of significance may be reduced as compared to a purely random design.

In conclusion, then, we feel that the split-litter procedure, when used with the realization that some genetic factors and some environmental factors are probably controlled while others are not, can be an efficient design. However, the experimenter should be aware of the shortcomings of this method and should not apply it blindly to all problems. In coming to a decision as to the advisability of using the split-litter procedure, the research worker has to decide whether the control gained over some (generally unknown) genetic factors and some constant environmental factors present within the litter more than compensate for the additional variability introduced by the probable segregation of recessive and infrequent dominant genes and for the influence of certain variable factors present within the litter environment. In addition he should also be aware of the possibility of obtaining biased estimates and loss of precision in his statistical analysis.

If the experimenter is critically concerned with controlling genetic variables, then we suggest either using isogenic strains of *S*s or instituting a breeding program to determine the genetic bases for the behaviors which he is studying. In any case, the split-litter technique does not guarantee an identical distribution of genetic sources of variance to the various treatment groups in the frequent instances of partially inbred lines of rats, dogs, or other experimental animals from small breeding colonies.

REFERENCES

1. BAYROFF, A. G. The experimental social behavior of animals. I. The effects of early isolation of white rats on their later reactions to other white rats as measured by two periods of free choices. *J. comp. Psychol.*, 1936, **21**, 67–81.
2. COREY, S. M. Equating groups in comparative experiments. *J. comp. Psychol.*, 1930, **10**, 287–294.
3. GINSBURG, B. E. Genetics and physiology of the nervous system. *Proc. Ass. Res. nerv. ment. Dis.*, 1954, **33**, 39–56.
4. HANSEN, M. H., HURWITZ, W. N., & MADOW, W. G. *Sample survey methods and theory*. Vol. I. New York: Wiley, 1953.
5. HEBB, D. O., & THOMPSON, W. R. The social significance of animal studies. In G. Lindzey (Ed.), *Handbook of social psychology*. Cambridge, Mass.: Addison-Wesley, 1954.
6. LI, CHING CHUN. *An introduction to population genetics*. Peiping, China: National Peking University Press, 1948.
7. LUSH, J. L. *Animal breeding plants* (2nd ed.). Ames, Iowa: Iowa State Coll. Press, 1943.
8. MUNN, N. L. *Handbook of psychological research on the rat*. New York: Houghton Mifflin, 1950.
9. RUSSELL, W. L. Inbred and hybrid animals and their value in research. In G. D. Snell (Ed.), *Biology of the laboratory mouse*. Philadelphia: Blakiston, 1941. Ch. 10, pp. 325–348.
10. SCOTT, J. P. Genetics as a tool in experimental psychological research. *Amer. Psychologist*, 1949, **4**, 526–530.
11. WRIGHT, S. Coefficients of inbreeding and relationship. *Amer. Natur.*, 1922, **56**, 330–338.
12. WRIGHT, S. Mendelian analysis of the pure breeds of livestock. I. The measurement of inbreeding and relationship. *J. Hered.*, 1923, **14**, 339–348.

Received June 1, 1956.

PSYCHOLOGICAL BULLETIN
Vol. 54, No. 2, 1957

ESTIMATING INTERACTION EFFECTS AMONG OVERLAPPING PAIRS

PHILIP J. RUNKEL
Bureau of Educational Research, University of Illinois

J. E. KEITH SMITH
Lincoln Laboratory, Massachusetts Institute of Technology

AND THEODORE M. NEWCOMB
University of Michigan

In studying communication and interaction among individuals, it frequently happens that the investigator observes some quantity associable with a pair of persons, but not observable in connection with either individual considered separately. Examples are attitude agreement between the two persons, amount of communication between them, belonging to the same family, or any relation which may be treated as symmetrical. When such binary relations are studied within groups of persons, it may occur that a person who is a member of one pair from which a measure is taken will also be a member of another pair from which a measure is taken. The experimenter then finds himself with measures which are not experimentally independent because the same individual contributed to both measures. This paper offers a method, given a collection of scores obtained by observing pairs of persons, of constructing scores in which the contributions of individuals are held constant so that the variability among the resulting "interaction scores" may be attributed to the conditions (e.g., communication or agreement) specifying the obtained pair-scores and not to characteristics associated with the persons individually. In regard to establishing a score for an observed unit such that the score is not biased by the fact that different units overlap the same individuals, this method may be considered a contribution to the same problem-area treated in recent papers by Luce, Macy, and Tagiuri (**2**), by Tagiuri, Bruner, and Kogan (**3**), and by Winer (**4**).

THE LINEAR HYPOTHESIS

The present method of constructing "interaction scores" from a collection of scores obtained from the observation of pairs of individuals rests on the hypothesis (cf. **1**, Ch. 5, 6) that the obtained pair-score y_{ij} consists of a linear combination of

μ = the population mean pair-score over all pairs,
b_i = the deviation from the mean due to person i,
b_j = the deviation from the mean due to person j, and
$(b)_{ij}$ = the deviation from $\mu+b_i+b_j$ due to the interaction of persons i and j. It is the estimation of this last component in which we are interested.

Accordingly, we begin by assuming that

$$y_{ij}=\mu+b_i+b_j+(b)_{ij}, \qquad [1]$$

and summing the pair-scores over the persons j, we have for any person i,

$$y_{i.}=\mu+b_i+b_.+(b)_{i.}. \qquad [2]$$

Now, since the population of persons is $\sum i=\sum j$, the sums of deviations from the mean are:

$$\sum_i b_i = \sum_j b_j = 0, \text{ and}$$

$$\sum_i (b)_{ij} = \sum_j (b)_{ij} = \sum_{i,j} (b)_{ij} = 0.$$

Then, summing the pair-scores containing person i, we have from Equation 2:

$$\sum_j y_{ij} = (N-1)\mu + (N-1)b_i + 0 + 0$$

where N is the number of persons, and $N-1$ the number of pairs which include person i. Or,

$$\frac{\sum_j y_{ij}}{N-1} = \mu + b_i.$$

And for all pairs, similarly, from Equation 1:

$$\frac{\sum_{i,j} y_{ij}}{\frac{N(N-1)}{2}} = \mu.$$

And for any sample, respectively, we have the expectations

$$E\left[\frac{\sum_j y_{ij}}{N-1}\right] = \mu + b_i, \text{ and} \qquad [3]$$

$$E\left[\frac{\sum_{i,j} y_{ij}}{\frac{N(N-1)}{2}}\right] = \mu. \qquad [4]$$

Thus from Equations 3 and 4, the estimated deviation score of person i is

$$\hat{b}_i = \frac{\sum_j y_{ij}}{N-1} - \frac{\sum_{i,j} y_{ij}}{N(N-1)/2}.$$

Now from Equation 1 we have

$$(b)_{ij} = y_{ij} - \mu - b_i - b_j,$$

and substituting the estimates we have

$$(\hat{b})_{ij} = y_{ij} - m_{ij} - \hat{b}_i - \hat{b}_j,$$

which with the appropriate substitutions becomes

$$(\hat{b})_{ij} = y_{ij} + \frac{\sum_{i,j} y_{ij}}{N(N-1)/2} - \frac{\sum_j y_{ij}}{N-1} - \frac{\sum_i y_{ij}}{N-1} \qquad [5]$$

for the estimated interaction between persons i and j.

Finally, the variance of the interaction within the pairs is

$$s^2_{(b)_{ij}} = \frac{\sum_{i,j} (\hat{b})^2_{ij}}{\frac{N(N-1)}{2}} - \left[\frac{\sum_{i,j} (\hat{b})_{ij}}{\frac{N(N-1)}{2}}\right]^2,$$

and since the last term is zero,

$$\frac{N(N-1)}{2} s^2 = \sum_{i,j} (\hat{b})^2_{ij}. \qquad [6]$$

Testing the Significance of the Difference Between Mean Interaction Effects

For the pair (i, j) in a group of $N(N-1)/2$ pairs, the estimated interaction effect on the pair-score is given by Equation 5. Now, since $\sum(\hat{b})_{ij} = 0$, the mean interaction effect on a *subset* of the $N(N-1)/2$ pair-scores will be significantly different from the mean effect on the remaining pair-scores if and only if the effect on the first-mentioned subset is significantly different from zero.

Therefore, the effect of a treatment on the pair-scores of a subgroup may be tested for its effect on the mean interaction level by computing the estimated interaction for each score

in the subgroup and testing by the t test whether the mean interaction is different from zero for that subgroup.

The data shown in Table 1 were collected during a methodological study conducted in the summer of 1953. Thirteen undergraduates at the University of Michigan were assigned randomly to pairs, and some pairs were given issues to discuss concerning the previous November's national election. Thirty-four of the possible 78 pairs were discussant pairs. Before and after each pair's discussion, the attitude of each member of the pair was measured in regard to possible consequences for various areas of national policy had Stevenson been elected. Pretest and posttest were five weeks apart. The discussion by the several pairs took place on different dates. Each pair's discussion was about 20 minutes long. A measure of attitude agreement was computed for each pair. Each figure in the body of Table 1 is an index of change in attitude agreement within a pair from pretest to posttest.

Each person number at the left of Table 1 labels a column and a row. The table is read as follows. The pair composed of persons 2 and 7 increased in agreement by an index of 2, persons 8 and 10 decreased in agreement by an index of 1, and so forth. The last column at the right shows the mean change in agreement for all pairs in which person i was a

TABLE 1

OBSERVED SCORES FOR CHANGE IN AGREEMENT AS TO CONSEQUENCES OF STEVENSON'S ELECTION AND MEAN SCORES FOR PAIRS CONTAINING PERSON i

Persons	Obtained Pair-scores												$\frac{\sum_i y_{ij}}{N-1}$
1	0	0	2	3	0	0	−4	−1	3	1	−2	1	0.250
2		−2	2	3	0	2	0	3	3	5	−2	−3	0.917
3			2	3	2	2	0	3	3	1	0	−1	1.083
4				1	6	−2	2	5	1	3	4	1	2.250
5					1	−3	1	2	−2	−2	3	0	0.833
6						−2	0	1	3	−1	4	3	1.417
7							−2	−3	−5	−5	0	1	−1.417
8								1	−1	1	0	5	0.250
9									0	−2	4	2	1.250
10										0	9	4	1.500
11											7	0	0.667
12												3	2.500
13													1.333

$$\frac{\sum_{i,j} y_{ij}}{N(N-1)/2} = 0.987$$

TABLE 2

COMPUTATION OF INTERACTION SCORES IN RESPECT TO CHANGE IN AGREEMENT ON THE PART OF DISCUSSANT PAIRS, FOR TESTING DIFFERENCE OF DISCUSSANT MEAN VERSUS NONDISCUSSANT MEAN

Pair i, j	y_{ij}	$\frac{\sum_j y_{ij}}{N-1}$	$\frac{\sum_i y_{ij}}{N-1}$	Interaction Score $(\hat{b})_{ij}$
1, 2	0	.250	.917	−.180
1, 3	0	.250	1.083	−.346
1, 5	3	.250	.833	2.904
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
9, 12	4	1.250	2.500	1.237
10, 12	9	1.500	2.500	5.987
11, 13	0	.667	1.333	−1.013
$n=34$			$\sum(\hat{b})_{ij}=$	26.561
			$\sum(\hat{b}^2_{ij})=$	164.728

member. The mean change for all pairs, shown to the left of the table, was 0.987.

Table 2 shows the method of computing the significance of the mean interaction effect of one of two treatments. The discussant pairs are listed in the first column of the table. Each line of the table represents the computation prescribed by Equation 5 for a particular pair i, j. The mean for the entire group of pairs, 0.987, is added to the obtained score for the pair, and from this sum are subtracted the means for each member of the pair. The resulting estimated interaction scores represent the changes in agreement which are not due to a tendency of the group as a whole to change in agreement, nor to a tendency peculiar to either of the individuals, but to the effect of the interaction (communication) between two particular individuals. These scores may next be subjected to the t test, and the result for the example of Tables 1 and 2 is $t=2.18$, which for 33 degrees of freedom and two tails is significant beyond the .05 level. Since this result argues that the mean interaction score for the discussant pairs is different from zero, and since the sum of all the interaction scores (both discussant and nondiscussant) must be zero, we can conclude that the mean interaction score for the discussant pairs is greater than the mean score for the nondiscussant pairs.

If fewer than all possible pairs enter into an experiment, the computation proceeds in the same way, but with the number of pairs in the denominators of the terms of Equation 5 reduced accordingly.

TESTING THE SIGNIFICANCE OF THE VARIANCE RATIO

Since the variance of a group of scores must be computed about the group's own mean, the computation of the F test using interaction scores proceeds a little differently from the test of the difference between means. The interaction scores are computed separately for the discussant group of pairs and for the nondiscussant group of pairs. Table 3 shows the observed scores for the discussant group with the means for each person, and Table 4 shows the same for the nondiscussant group of pairs. Tables 3 and 4 are read as follows. The cells with entries in Table 3 indicate the discussant pairs; those in Table 4 indicate the nondiscussant pairs. (The cells with entries in Table 3 are the empty cells in Table 4, and conversely.) The means for person i in the right-hand column are computed by dividing the sum of the entries for person i by the number of entries, which latter is symbolized in the column-heading by n_i. In the expression for the group mean at the left of this table, n_1 stands for the

TABLE 3

OBSERVED SCORES FOR CHANGE IN AGREEMENT AS TO CONSEQUENCES OF STEVENSON'S ELECTION, AND MEAN SCORES FOR PAIR CONTAINING PERSON i, FOR DISCUSSANT PAIRS ONLY

	Obtained Pair-scores												$\frac{\sum_j y_{ij}}{n_i}$
1	0	0		3		0		−1	3		−2		0.429
	2		2	3		2	0						1.400
		3		3	2	2		3	3			−1	1.713
			4		6			5				1	3.500
				5	1	−3					3		1.667
					6		0	1			4		2.333
		Persons				7					0		0.250
							8	1		1		5	1.400
								9	0		4		1.856
									10		9		3.750
										11		0	0.500
											12		3.000
$\frac{\sum_{i,j} y_{ij}}{n_1} = 1.765$												13	1.250

number of discussant pairs (Table 3) and n_2 for the number of nondiscussant pairs (Table 4).

Tables 5 and 6 show the computation of interaction scores. Each row in these tables is again a solution of Equation 5, except that the mean indicated by the second term on the right-hand side of Equation 5 is now the mean of the particular group of pairs being examined.

For each group of pairs we can adapt Equation 6. The sum of squares for the first group will be

$$n_1 s^2_1 = \sum_{i,j} (\hat{b})^2_{ij}$$

with $n_1 - 1$ degrees of freedom, and similarly for the second group with $n_2 - 1$ degrees of freedom. For our example, it turns out that the second group has the larger mean square, and we have

$$F = \frac{\frac{\sum_{i,j} (\hat{b}_2)^2_{ij}}{(n_2 - 1)}}{\frac{\sum_{i,j} (\hat{b}_1)^2_{ij}}{(n_1 - 1)}} = \frac{\frac{183.48}{43}}{\frac{125.28}{33}} = 1.12.$$

This result, nonsignificant, indicates that the significance of the difference in mean pair-agreement between the two groups resides in the relative level of interaction effect, and not in the variability of the interaction scores.

TABLE 4

OBSERVED SCORES FOR CHANGE IN AGREEMENT AS TO CONSEQUENCES OF STEVENSON'S ELECTION, AND MEAN SCORES FOR PAIRS CONTAINING PERSON i, FOR NONDISCUSSANT PAIRS ONLY

Persons	Obtained Pair-scores												$\frac{\sum_j y_{ij}}{n_i}$
1			2		0		−4			1		1	0.000
2	−2				0			3	3	5	−2	−3	0.571
3	2						0			1	0		0.200
4	1				−2		2		1	3	4		1.625
5							1	2	−2	−2		0	0.000
6	−2								3	−1		3	0.500
7	−2		−3		−5		−5					1	−2.571
8					−1					0			−0.571
9					−2					2			0.400
10	0									4			0.375
11	7												0.700
12	3												2.000
13													1.375

$\frac{\sum_{i,j} y_{ij}}{n_2} = 0.386$

TABLE 5

COMPUTATION OF INTERACTION SCORES IN RESPECT TO CHANGE IN AGREEMENT ON THE PART OF DISCUSSANT PAIRS, FOR TESTING THE VARIANCE RATIO

Pair i, j	y_{ij}	$\frac{\sum_j y_{ij}}{n_i}$	$\frac{\sum_i y_{ij}}{n_j}$	Interaction Score $(\hat{b})_{ij}$
1, 2	0	.429	1.400	−.064
1, 3	0	.429	1.713	−.377
1, 5	3	.429	1.667	2.669
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
9, 12	4	1.856	3.000	.909
10, 12	9	3.750	3.000	4.015
11, 13	0	.500	1.250	.015
$n_1 = 34$			$\sum (\hat{b})_{ij} = 0$	
			$\sum (\hat{b})^2_{ij} = 125.280$	

TABLE 6

COMPUTATION OF INTERACTION SCORES IN RESPECT TO CHANGE IN AGREEMENT ON THE PART OF NONDISCUSSANT PAIRS, FOR TESTING THE VARIANCE RATIO

Pair i, j	y_{ij}	$\frac{\sum_j y_{ij}}{n_i}$	$\frac{\sum_i y_{ij}}{n_j}$	Interaction Score $(\hat{b})_{ij}$
1, 4	2	0	1.625	.761
1, 6	0	0	.500	−.114
1, 8	−4	0	−.571	−3.043
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
10, 13	4	.375	1.375	2.636
11, 12	7	.700	2.000	4.686
12, 13	3	2.000	1.375	.011
$n_2 = 44$			$\sum (\hat{b})_{ij} = 0$	
			$\sum (\hat{b})^2_{ij} = 183.482$	

SUMMARY

A method of computing the interaction effect of variables measured by observing interacting pairs of persons has been presented. The estimate of the interaction effect utilizes the linear hypothesis. This technique was developed for use in situations where one person may be a member of more than one of the pairs being studied. It can be used wherever measurements associated with pairs are taken and where the experimenter's interest is in the effects of interaction within the pair. Only some of all possible pairs of the subjects need be observed.

REFERENCES

1. KEMPTHORNE, O. *The design and analysis of experiments.* New York: Wiley, 1952.
2. LUCE, R. D., MACY, J., and TAGIURI, R. A statistical model for relational analysis. *Psychometrika,* 1955, **20**, 319–327.
3. TAGIURI, R., BRUNER, J. S., & KOGAN, N. Estimating the chance expectancies of diadic relationships within a group. *Psychol. Bull.,* 1955, **52**, 122–131.
4. WINER, B. J. A measure of interrelationship for overlapping groups. *Psychometrika,* 1955, **20**, 63–68.

Received June 11, 1956.

PSYCHOLOGICAL BULLETIN
Vol. 54, No. 2, 1957

AN ADDITION TO SCHAEFFER AND LEVITT'S "KENDALL'S TAU"

MARSHALL B. JONES[1]
U. S. Naval School of Aviation Medicine, Pensacola, Florida

In their recent review of the literature concerning Kendall's tau Schaeffer and Levitt (**3**) remark that "generally applicable tests of the significance of any partial tau are not yet available." While this assertion is true of the partial rank correlation coefficient originally described by Kendall (**2**), a partial tau, applicable in situations where the variable whose effects are to be removed is nominal (**4**, p. 25), has been described and a test of significance appropriate to the hypothesis of zero partial correlation has been developed. The statistic in question was first described by the writer (**1**) in 1954. In the same paper the mean and the variance of the sampling distribution were obtained; and the limit distribution was proved to be normal. In 1955, and independently, Torgerson (**5**) described the same statistic. In addition to the results already described Torgerson offered a correction for continuity and a discussion of tied ranks. Very recently, Torgerson republished his original interoffice draft in *Psychometrika* (**6**). The purpose of this note is to append a brief description of the statistic in question to Schaeffer and Levitt's review.

Suppose we wish to correlate two variables, X and Y, partialling out the effects of a nominal variable Z. For purposes of illustration let Z be geographical region. The data might appear as below

	N	S	E	W
Y	26	29	55	14
	32	71	43	15
	45	63		93
		82		

[1]

where the numbers are the raw scores on X arranged (from top to bottom) for increasing values on the paired Y score. Ranking the scores in each column we have

	N	S	E	W
Y	1	1	2	1
	2	3	1	2
	3	2		3
		4		

[2]

$V_n = 3 \quad V_s = 5 \quad V_e = 0 \quad V_w = 3.$

Let V_i count the number of times a higher rank on X_i precedes (comes higher in the ith column than) a lower rank. Letting

$$V = \sum_{i=1}^{k} V_i,$$

V varies between

$$M = \frac{\sum_{i=1}^{k} (n_i^2 - n_i)}{2} \qquad [3]$$

and 0, where n_i represents the number of cases in the ith column. The

[1] Opinions or conclusions contained in this note are those of the author. They are not to be construed as necessarily reflecting the view or the endorsement of the Navy Department.

partial rank correlation coefficient is then defined as

$$\frac{2V-M}{M}. \qquad [4]$$

In this description we will consider only the case of no ties on either X or Y. For a discussion of tied ranks, see Torgerson (6, p. 151).

Our concern is with the sampling distribution of V under the hypothesis that any one of the $n_1! \cdots n_k!$ possible patterns in the k columns is as likely as any other. The distribution has a mean equal to $M/2$ and a standard deviation

$$\sigma_V = \left[\frac{\sum_{i=1}^{k} n_i(n_i-1)(2n_i+5)}{72} \right]^{1/2}. \qquad [5]$$

The distribution tends rapidly to normality, even with few columns and few cases within the columns (6, p. 147). Therefore, to test for significance we need only calculate

$$\frac{V-\frac{M}{2}}{\sigma_V} \qquad [6]$$

and refer the result to tables of the normal curve.

REFERENCES

1. JONES, M. B. Partial rank correlation: a special case. *U. S. Naval Sch. Aviat. Med.*, 20 October 1954, Rep. No. NM 001 058.23.
2. KENDALL, M. G. *Rank correlation methods.* London: Griffin, 1948.
3. SCHAEFFER, M. S., & LEVITT, E. E. Concerning Kendall's Tau, a nonparametric correlation coefficient. *Psychol. Bull.*, 1956, **53**, 338–346.
4. STEVENS, S. S. (Ed.) *Handbook of experimental psychology.* New York: Wiley, 1951.
5. TORGERSON, W. S. Note on a nonparametric test for correlation for data expressed in the form of rank orders within subgroups. RB-55-3. Princeton: *Educational Testing Service*, February 1955.
6. TORGERSON, W. S. A nonparametric test of correlation using rank orders within subgroups. *Psychometrika*, 1956, **21**, 145–152.

Received July 17, 1956.